Stochastic Dynamics: Mathematical Foundations of Diffusion and Score-Based Models

Rehan Ahmad

Department of Physics, LUMS

Contents

1 Introduction

1.1 Motivation

One night, as I lay awake in my bed after another fruitless day of reading too many papers on diffusion models while not making much progress, a strange thing happened. Rather than being irritated as I would often be after the task, I found myself transfixed by my surroundings. I thought of how everything around my house—*the fan , the closets, the books*— from the huge to the minuscule, could gradually be brought into existence from random noise. Diffusion models stands in a long list of generative models such as GANs, VQ-VAEs and Energy-based Models. Yet, there is a certain elegance to their ability of reversing the arrow of time to create something out of nothing. Coming from a background in physics, I find it fascinating that the random motion of particles diffusing through a medium can evolve into a sophisticated framework for generating data.

In this Independent Study, we are concerned with the stochastic processes underlying the current state-of-the-art (SoTA) score-based and diffusion models from a physicist's perspective. The goal is to build a deep understanding of these powerful models work from scratch. Starting with the fundamentals, we will explore Stochastic Differential Equations (SDEs) which form the backbone of these models before delving into key concepts in generative modeling such as the Score Function, Variance Preserving SDEs, and Variance Exploding SDEs. This study assumes a working familiarity with multivariate calculus, linear algebra, ordinary differential equations, probability, and statistical mechanics. All other concepts will be developed from the ground up. Although this work will remain an open book for some time, the ultimate goal is to produce a self-contained reference that explains the underlying framework of diffusion models in a way that resonates with someone from a physics background. My hope is that it will be as rewarding to explore as studying classical mechanics or other foundational subjects in physics—offering a reason for studying the subject for a sense of elegance and discovery rather than any notion of utility.

2 Introduction to Stochastic Differential Equations

2.1 From ODEs to SDEs

Imagine you are modeling the growth of a population $x(t)$ of bacteria. In a deterministic setting, we might use the following differential equation to model the growth of the population:

$$\frac{dx}{dt} = rx \tag{2.1}$$

where r is the growth rate. This equation tells us that the rate of change of the population is proportional to its current size. Given an initial population $x(0)$, the future population at any time t is completely determined. The solution to this ODE is trivial and is taught in any first course:

$$x(t) = x(0)e^{rt}. \tag{2.2}$$

Now, suppose the growth rate r is not constant but fluctuates due to random environmental factors such as changes in nutrient availability or temperature. In this case, the change in x is no longer deterministic. This is the subject of stochastic differential equations (SDEs).

To get a flavor of SDEs, let us begin with a linear ordinary differential equation with a deterministic driving force that describes a damped harmonic oscillator under an external force $f(t)$:

$$\frac{dx}{dt} = -\gamma x + f(t). \tag{2.3}$$

We can discretize this ODE and write it in terms of differentials:

$$\Delta x(t_n) = x(t_n)\Delta t + f(t_n)\Delta t \tag{2.4}$$

where $f(t_n)\Delta t$ is the driving term. The value of $x(t_n + \Delta t)$ is then given by:

$$x(t_n + \Delta t) = x(t_n) + \Delta x(t_n)$$

Substituting $\Delta x(t_n)$ from (2.4), we have:

$$x(t_n + \Delta t) = x(t_n) + x(t_n)\Delta t + f(t_n)\Delta t. \tag{2.5}$$

If we know the value of x at $t = 0$, then:

$$x(\Delta t) = x(0)(1 + \Delta t) + f(0)\Delta t. \tag{2.6}$$

What we are particularly interested in is the scenario where the driving term $f(t_n)\Delta t$ becomes random at each time step t_n. This involves replacing $f(t_n)$ with a random variable y_n at each t_n. As a result, the difference equation (2.4) transforms

into the following:

$$\Delta x(t_n) = x(t_n)\Delta t + y_n \Delta t \tag{2.7}$$

where $y_n \sim P(Y = y_1, \ldots, y_i)$.

This is called a *stochastic difference equation*. It states that at each time t_n, we pick a value for the random variable y_n sampled from its probability density $P(Y)$ and add $y_n \Delta t$ to $x(t_n)$. Consequently, we can no longer predict the exact value of x at some future time T in advance. Instead, $x(T)$ depends on the cumulative effect of all random increments y_n up to time T, which are determined sequentially as the process evolves.

The solution for x at time Δt is:

$$x(\Delta t) = x(0)(1 + \Delta t) + y_0 \Delta t \tag{2.8}$$

Thus, $x(\Delta t)$ is now a random variable. If the initial condition $x(0)$ is fixed (i.e., not random), then $x(t)$ can be expressed as a linear transformation of the random variable y_0, which represents the noise increment over the interval $[0, \Delta t]$. On the other hand, if $x(0)$ is also a random variable, then $x(t)$ is a linear combination of two random variables: the initial condition $x(0)$ and the noise increment y_0.

When we proceed to the next time step to calculate $x(2\Delta t)$, this value depends on $x(0)$, y_0, and the noise increment for the second step y_1. At each time step, the solution of the stochastic difference equation $x(t_n)$ is a random variable, and this random variable evolves as time progresses. Thus, solving a stochastic difference equation requires determining the probability density of $x(t_n)$ for all future times t_n. This process involves deriving the probability density of $x(t_n)$ from:

- The probability densities of the noise increments y_n.
- The probability density of $x(0)$, if $x(0)$ is random.

Stochastic differential equations (SDEs) are obtained by taking the limit $\Delta t \to 0$ in stochastic difference equations. In this continuous-time framework, the solution to an SDE is characterized not by a single trajectory but by a probability density function that describes the value of x at future times t. Just as with ordinary (deterministic) differential equations, finding a closed-form solution to an SDE is not always possible. However, in many simple cases, explicit solutions can be derived. For more complex systems, numerical methods or approximations are typically employed to study the evolution of the probability density over time.

In addition to obtaining the probability density for x at future times t_n, we can also ask how x evolves with time given a specific set of values for the random increments y_n.

This focuses on two key aspects:

- The probability density of x at specific times.
- The sample paths of x, which describe specific realizations of its evolution under different noise realizations.

Realization of Noise

Let y_i represent the random noise increments over discrete time intervals $[\Delta t, 2\Delta t, \dots]$. A realization of the noise is a specific set of sampled values $\{y_1, y_2, \dots, y_n\}$ drawn from the noise's probability density $y_n \sim N(0, \Delta t)$

Sample Path of $x(t)$

A sample path is a trajectory of $x(t)$ over time that corresponds to a specific realization of the noise. The full solution to the SDE is the collection of all possible sample paths. This full solution is rarely needed in practice. Instead, we focus on:

- The marginal probability density $P(x, t)$: The probability density of $x(t)$ at each time t, obtained from the Fokker-Planck equation associated with the SDE.
- Correlation Functions: Quantities like $\mathbb{E}[x(t)x(t')]$ which describe how $x(t)$ at one time is statistically related to $x(t')$ at another.

2.2 Introduction to Weiner Increments

By "Gaussian noise," we mean that each of the random increments $y_0\Delta t$ has a Gaussian probability density. Specifically, $y_n \sim \mathcal{N}(0, \sigma^2)$. First, consider the simplest stochastic difference equation where the increment of x consists solely of the random increment $y_n\Delta t$. This reduces equation (2.7) to:

$$\Delta x(t_n) = y_n \Delta t \tag{2.9}$$

In literature, Gaussian noise is often referred to as ***Wiener noise***, and the random increment is expressed as:

$$\Delta W_n = y_n \Delta t \tag{2.10}$$

The discrete differential equation for x can thus be written as:

$$\Delta x(t_n) = \Delta W_n \tag{2.11}$$

Each Wiener increment ΔW_n is independent of the others and has the same probability density:

$$P(\Delta W) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\Delta W)^2}{2\sigma^2}} \tag{2.12}$$

This density is Gaussian with zero mean, and we set the variance as $\sigma^2 = V = \Delta t$. This choice is critical and as we will see in the next section, any other choice would

lead to unphysical systems For simplicity, we often denote a Wiener increment in a given time step Δt as ΔW, without referencing the subscript n, since all increments share the same distribution and are independent.

We can solve the difference equation (2.11) for x by starting with $x(0) = 0$ and repeatedly adding Δx. The solution is:

$$x_n \equiv x(n\Delta t) = \sum_{i=1}^{n-1} \Delta W_i \tag{2.13}$$

The probability density of x_n can now be calculated. Since the sum of Gaussian random variables is also Gaussian, the probability density of x_n is Gaussian. The mean and variance of x_n are the sums of the means and variances of ΔW_i, respectively, because the ΔW_i are independent. Thus:

$$\mu = \langle x_n \rangle = 0 \tag{2.14}$$

$$V(x_n) = n\sigma^2 = n\Delta t \tag{2.15}$$

The solution to the difference equation is then:

$$P(x_n) = \frac{1}{\sqrt{2\pi V}} e^{-\frac{x_n^2}{2V}} = \frac{1}{\sqrt{2\pi n\Delta t}} e^{-\frac{x_n^2}{2n\Delta t}} \tag{2.16}$$

We can easily cast the *difference equation* (2.11) *to a differential equation*:

$$dx = dW \tag{2.17}$$

To solve this, consider T future steps with N discrete time steps and take the limit as $N \to \infty$:

$$x(T) = \lim_{N\to\infty} \sum_{i=0}^{N-1} \Delta W_i = \int_0^T dW(t) = W(T) \tag{2.18}$$

Here, the stochastic integral $W(T) = \int_0^T dW(t)$ is defined as the *limit of the sum of all increments of the Wiener process.* For $x(T)$, the probability density is Gaussian because it sums independent Gaussian variables. The mean is zero since each variable has zero mean. The variance of $x(T)$ does not require the limit $N \to \infty$ because N factors out:

$$V(x(T)) = \sum_{i=1}^{N-1} V[\Delta W_i] = \sum_{i=1}^{N-1} \Delta t = N\Delta t = N(T/N) = T \tag{2.19}$$

Thus, the probability density of $W(T)$ is:

$$P(W(T)) = P(x(T)) = P(x, T) = \frac{1}{\sqrt{2\pi T}} e^{-\frac{x^2}{2T}} \tag{2.20}$$

Definition of Wiener Process

The Wiener process is defined as $W(T)$, while dW represents an increment of the Wiener process:

$$W(T) = \int_0^T dW \tag{2.21}$$

Note: While we often refer to dW as the Wiener process, the Wiener process is technically $W(T)$, and dW is its increment.

The variance of a wiener increment must satisfy Δt

In the derivation above, we assumed that the variance of Wiener increments is Δt which led to the fact that the variance of $x(T)$ is proportional to T. It turns out that setting the variance to any other value leads to an unphysical system. To show this, we set $V[\Delta W(\Delta t)] = \Delta t^\alpha$ and calculate the variance of $x(T)$ once again:

$$V(x(T)) = \sum_{i=1}^{N-1} V[\Delta W_i] = N(\Delta t)^\alpha = N\left(\frac{T}{N}\right)^\alpha = N^{(1-\alpha)}T^\alpha \tag{2.22}$$

Now we take the continuum limit $N \to \infty$ to obtain a stochastic differential equation. When $\alpha > 1$, we have:

$$\lim_{N\to\infty} V(x(T)) = T^\alpha \lim_{N\to\infty} N^{1-\alpha} = 0 \tag{2.23}$$

And when $\alpha < 1$, we have:

$$\lim_{N\to\infty} V(x(T)) = T^\alpha \lim_{N\to\infty} N^{1-\alpha} = \infty \tag{2.24}$$

Neither of these results make sense for obtaining a stochastic differential equation that describes real systems driven by noise. Thus, we are forced to choose $\alpha = 1$ and hence $V[\Delta W(\Delta t)] \propto \Delta t$.

2.3 General SDE Form and an Example of SDE

Given that we have introduced the fundamentals of SDEs, let us now introduce a bit of formalism. A general SDE for a single variable X_t can be written as:

$$dX_t = \underbrace{\mu(X_t, t)}_{\text{Drift term}}\, dt + \underbrace{\sigma(X_t, t)}_{\text{Diffusion term}}\, dW_t \tag{2.25}$$

Since the variance of dW_t must be proportional to dt, and any constant of propor-

tionality can always be absorbed into $\sigma(X_t, t)$, the variance of dW_t is defined to be equal to dt. Therefore, the probability density of dW_t is:

$$P(dW) = \frac{1}{\sqrt{2\pi dt}} e^{-\frac{(dW)^2}{2dt}} \tag{2.26}$$

Let us consider a simple SDE of the form:

$$dX = e^t dW \tag{2.27}$$

Where $W(t)$ is a Wiener process. This SDE can be expressed as:

$$dX = f(t)dW \tag{2.28}$$

To solve this equation, we integrate both sides from 0 to t:

$$X(t) - X(0) = \int_0^t e^s dW(s) \tag{2.29}$$

Assuming the initial condition $X(0) = 0$, we have:

$$X(t) = \int_0^t e^s dW(s) \tag{2.30}$$

This integral represents the accumulation of the stochastic process e^s weighted by the increments of the Wiener process over $[0, t]$.

Mean of $X(t)$

Since the integrand e^s and the Wiener process increments $dW(s)$ are independent and the expectation of $dW(s)$ is zero:

$$\mathbb{E}[X(t)] = \mathbb{E}\left[\int_0^t e^s dW(s)\right] = 0 \tag{2.31}$$

Variance of $X(t)$

The variance of $X(t)$ can be calculated using Itô isometry (a result we will prove later) :

$$V(X(t)) = \mathbb{E}\left[\left(\int_0^t e^s dW(s)\right)^2\right] = \int_0^t (e^s)^2 ds = \int_0^t e^{2s} ds \tag{2.32}$$

Evaluating the integral, we obtain:

$$V(X(t)) = \frac{e^{2t} - 1}{2} \tag{2.33}$$

3 Ito Calculus

3.1 Ito's Rule

Consider the solution to the simple differential equation:

$$dx = -\theta x \, dt \tag{3.1}$$

The solution to this equation is straightforward. The deterministic nature of this equation ensures a well-defined trajectory for $x(t)$. However, when we introduce stochastic elements, the rules for solving such equations become less straightforward, as we will see.

Deterministic Solution to (3.1)

Let us begin by revisiting how we find the solution to (3.1). The value of x at time $t + dt$ is the value at time t plus the infinitesimal change dx:

$$x(t + dt) = x(t) - \theta x(t) dt = (1 - \theta dt) x(t) \tag{3.2}$$

From Taylor series, $e^{\theta dt} \approx 1 + \theta dt$. We can therefore write the equation for $x(t + dt)$ as:

$$x(t + dt) = e^{-\theta dt} x(t) \tag{3.3}$$

This tells us that to move x from time t to $t + dt$, we merely have to multiply $x(t)$ by the factor $e^{-\theta dt}$. So to move by two lots of dt, we simply multiply by this

factor twice:

$$x(t+2dt) = e^{-\theta dt}x(t+dt) = e^{-\theta dt}[e^{-\theta dt}x(t)] = e^{-\theta\cdot 2dt}x(t). \tag{3.4}$$

To obtain $x(t+\tau)$, all we have to do is apply this relation repeatedly. Let us say that $dt = \tau/N$ for N as large as we want. Thus, dt is a small but finite time-step, and we can make it as small as we want. That means that to evolve x from time t to $t+\tau$, we can apply (3.3) N times:

$$x(t+\tau) = \left(e^{-\theta dt}\right)^N x(t) = e^{-\theta\sum_{n=1}^{N} dt}x(t) = e^{-\theta N dt}x(t) = e^{-\theta\tau}x(t) \tag{3.5}$$

is the solution to the differential equation. If θ is a function of time, so that the equation becomes:

$$dx = -\theta(t)x\,dt \tag{3.6}$$

then the line of attack remains pretty much the same. As before, we set $dt = \tau/N$ so that it is a small finite-time step, but this time we have to explicitly take the limit as $N \to \infty$ to obtain the solution to the differential equation:

$$\begin{aligned} x(t+\tau) &= \lim_{N\to\infty}\prod_{n=1}^{N}\left(e^{-\theta(t+ndt)dt}\right)x(t) \\ &= \lim_{N\to\infty} e^{-\sum_{n=1}^{N}\theta(t+ndt)dt}x(t) \\ &= e^{-\int_t^{t+\tau}\theta(t)dt}x(t) \end{aligned} \tag{3.7}$$

Notice that we were able to solve the equation (3.3) using $e^{\alpha dt} \approx 1+\alpha dt$. The approximation $e^{\alpha dt} \approx 1+\alpha dt$ works because the terms in the power series expansion for $e^{\alpha dt}$ that are second-order or higher in dt ($dt^2, dt^3, \ldots$) will vanish in comparison to dt as $dt \to 0$. The result of being able to ignore terms that are second-order and higher in the infinitesimal increment leads to the usual rules for differential equations. (It also means that any equation we write in terms of differentials dx and dt can alternatively be written in terms of derivatives.)

We will now spend some time showing that as $dt \to 0$, $(dW)^2$ does not tend to 0. We must therefore learn a new rule for the manipulation of stochastic differential equations as the second-order moments are no longer negligible. .

To examine whether $(dW)^2$ makes a non-zero contribution to the solution, we will sum $(dW)^2$ over all the time-steps for a finite time T. To do this, we will return to the discrete description of differentials:

$$\Delta X = \lim_{N\to\infty} \sum_{i=1}^{N} (dW_i)^2 \tag{3.8}$$

The first thing to note is that the expectation value of $(\Delta W)^2$ is equal to the variance of ΔW, because $\langle \Delta W \rangle = 0$. From $\text{Var}(\Delta W) = \langle (\Delta W)^2 \rangle - \langle \Delta W \rangle^2$, we have:

$$\langle \Delta W^2 \rangle = \Delta t \tag{3.9}$$

This tells us immediately that the expectation value of $(\Delta W)^2$ does not vanish with respect to the time-step Δt, and so the sum of these increments will not vanish when we sum over all the time-steps and take the infinitesimal limit (because we would have a Gaussian on our hands with a variance Δt). In fact, the expectation value of the sum of all the increments $(dW)^2$ from 0 to T is simply T:

$$\langle \int_0^T (dW)^2 \rangle = \int_0^T \langle (dW)^2 \rangle = \int_0^T dt = T \tag{3.10}$$

Claim: The Integral of $(\Delta W)^2$ is Deterministic

In order to show that the integral of $(\Delta W)^2$ is deterministic, we will look at the variance $\int_0^T (dW)^2$ and show that it is zero. To do this, we will first argue about the differential $(\Delta W)^2$ itself. We found that the expectation of $(\Delta W)^2$ is proportional to Δt, and so we have a strong hunch that the variance of $(\Delta W)^2$ must be proportional to $(\Delta t)^2$. *Hint:* Think in terms of $\text{Var}(x) = \mathbb{E}[x^2] - (\mathbb{E}[x])^2$. With this intuition in mind, we calculate the variance of $(\Delta W)^2$ below.

Proof

To begin with, consider the probability density for ΔW:

$$P(\Delta W) = \frac{1}{\sqrt{2\pi\Delta t}} e^{-\frac{(\Delta W)^2}{2\Delta t}} \tag{3.11}$$

Using the change of variable theorem we can find the probability distribution for $(\Delta W)^2$. We define $z = (\Delta W)^2$. For this, we have two cases: $\Delta W = \sqrt{z}$ and $\Delta W = -\sqrt{z}$. The absolute value of the derivative $\frac{d\Delta W}{dz}$ for $z = (\Delta W)^2$ is:

$$\left| \frac{d\Delta W}{dz} \right| = \left| \frac{d}{dz} (\pm\sqrt{z}) \right| = \frac{1}{2\sqrt{z}}.$$

Using the change of variables formula for PDFs, we have:

$$f_z(z) = f_{\Delta W}(\sqrt{z}) \left| \frac{d\Delta W}{dz} \right| + f_{\Delta W}(-\sqrt{z}) \left| \frac{d\Delta W}{dz} \right|.$$

Since $f_{\Delta W}(x)$ is even, $f_{\Delta W}(-\Delta W) = f_{\Delta W}(\Delta W)$. Thus:

$$f_z(z) = \frac{f_{\Delta W}(\sqrt{z})}{\sqrt{z}}.$$

Substituting the original PDF, we obtain:

$$f_z((\Delta W)^2) = \frac{1}{\sqrt{2\pi \Delta t z}} e^{-\frac{z}{2\Delta t}}. \tag{3.12}$$

This is a **chi-squared distribution** with one degree of freedom, scaled by a factor of $2\Delta t$. The variance of the distribution is:

$$V((\Delta W)^2) = 2(\Delta t)^2 \tag{3.13}$$

The variance of the sum of all the $(\Delta W)^2$ is:

$$V\left[\sum_{n=1}^{N-1} (\Delta W)^2\right] = \sum_{n=1}^{N-1} V((\Delta W)^2) = \sum_{n=1}^{N-1} 2(\Delta t)^2 = 2N \left(\frac{T}{N}\right)^2 = \frac{2T^2}{N} \tag{3.14}$$

Finally, we can perform what we set out to do, namely evaluating the integral in equation (3.8):

$$\lim_{N\to\infty} \left[\sum_{n=1}^{N-1} V((\Delta W)^2)\right] = \lim_{N\to\infty} \frac{2T^2}{N} = 0. \tag{3.15}$$

Since the integral of all the $(dW)^2$ is deterministic, it is equal to its mean T. That is:

$$\int_0^T (dW)^2 = T = \int_0^T dt \tag{3.16}$$

Thus, we have the surprising result:

$$dW^2 = dt \tag{3.17}$$

This is officially known as ***Ito's rule***. It is the fundamental rule for solving stochastic differential equations that contain Gaussian noise.

3.2 Ito's Lemma

Due to the Ito rule established in equation (3.17), we will have to keep all terms that are first order in dt and dW, as well as all terms that are second order in dW. In fact, wherever we find terms that are second order in dW, we can simply replace them with dt.

To see how this works, consider a simple example in which we want to know the differential equation for:

$$y = x^2 \tag{3.18}$$

The differential equation assumes the form:

$$\begin{aligned} dy &= y(t+dt) - y(t) \\ &= x(t+dt)^2 - x(t)^2 \\ &= (x+dx)^2 - x^2 \\ &= x^2 + 2xdx + (dx)^2 - x^2 \\ &= 2xdx + (dx)^2 \end{aligned} \tag{3.19}$$

Had x been deterministic, $(dx)^2$ would vanish in the continuum limit.
We would then have by the usual rule of calculus:

$$dy = 2xdx \quad \text{or} \quad \frac{dy}{dx} = 2x \tag{3.20}$$

However, if X is a random variable obeying the stochastic differential equation 2.25, then the equation (3.19) becomes the following:

$$\begin{aligned} dY &= 2xdX + (dX)^2 \\ &= 2x(fdt + gdW) + (fdt + gdW)^2 \\ &= 2x(fdt + gdW) + f^2dt^2 + g^2dW^2 + 2fgdtdW \end{aligned} \tag{3.21}$$

Just like in normal calculus, we can ignore cross-differentials like $dt\,dW$ and dt^2:

$$dY = 2fXdt + 2xgdW + g^2dW^2. \tag{3.22}$$

Using Ito's rule:

$$dY = (2fX + g^2)dt + 2xgdW \tag{3.23}$$

Ito's Lemma

There is a simple way to calculate the increment of any nonlinear function $y(X)$ in terms of the first and second powers of the increment of X, where X is a random variable in comparison to the derivation above. All we have to do is use the Taylor series expansion for $y(X)$, truncated at the second term:

$$dy(X) = \frac{dy}{dX}dX + \frac{1}{2}\frac{d^2y}{dX^2}(dX)^2 \tag{3.24}$$

If y is also an explicit function of time as well as X, then this becomes:

$$dy(t, X) = \frac{\partial y}{\partial X}dX + \frac{\partial y}{\partial t}dt + \frac{1}{2}\frac{\partial^2 y}{\partial X^2}(dX)^2 \tag{3.25}$$

3.3 Ornstein-Uhlenbeck Process

The Ornstein-Uhlenbeck (OU) process is a type of stochastic process used to model mean-reverting behavior. It was introduced by Leonard Ornstein and George Eugene Uhlenbeck in 1930 to describe the velocity of a particle undergoing Brownian motion under the influence of friction.

Mean-Reverting Behavior

Imagine a particle diffusing in a liquid where there is some friction or restoring force that pulls the particle back towards a central point (mean position) rather than allowing it to drift indefinitely. In this context, the position X_t of the particle at time t can be described by a stochastic differential equation governed by two terms, $\theta(\mu - X_t)$ and σdW_t:

$$dX_t = \theta(\mu - X_t)dt + \sigma dW_t \tag{3.26}$$

In this equation, the important thing to take note of is the **drift term**. θ acts as a restoring force, causing the particle to revert to the mean position μ. To understand how this would play out, suppose the particle starts at the position X_t far from the mean μ. For example, if μ is at the origin (0) and the particle is initially at $X_0 = 5$, the term $\theta(\mu - X_t)dt$ acts as a restoring force that pulls the particle back towards the mean position μ. The larger θ is, the stronger the pull back to μ. Despite the mean-reverting force, the particle is subject to random fluctuations due to the noise term σdW_t. This represents the random kicks from the surrounding fluid molecules.

Visually, processes obeying (3.26) look like the following:

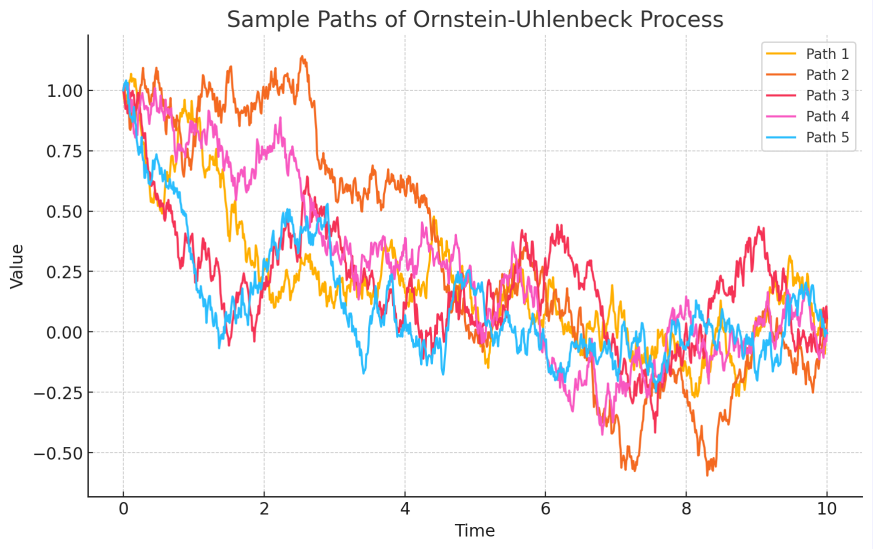


Figure 1: Sample Paths obeyings the Orstein-Uhlenbeck Equation

Here we can see that even though two of the three sample paths start far away from the mean, they quickly converge to a region around the mean. Once in the vicinity of the mean ($\mu = 0$, they move about it in arcs through the momentum factor.

In order to solve 3.26, we define a new random variable Y_t:

$$Y_t = X_t - \mu \tag{3.27}$$

The differential element of dY_t is given by:

$$\begin{aligned} dY_t &= dX_t \\ &= \theta(\mu - X_t)dt + \sigma dW_t \\ &= -\theta(X_t - \mu)dt + \sigma dW_t \\ &= -\theta Y_t dt + \sigma dW_t. \end{aligned} \tag{3.28}$$

Equation 3.28 is the reason why we say that Orstein-Uhlenbeck equation is governed by "additive noise". The term "additive noise" refers to the fact that **the noise does not itself depend on** Y_t, but is merely added to any other terms that appear in the equation for dx. This equation is called the Ornstein–Uhlenbeck equation, and its solution is called the Ornstein–Uhlenbeck process. The next step is to recognize that

we are equating the derivative of a random variable with itself:

$$\begin{aligned} dY_t &\propto \theta Y_t dt \\ Y_t &\propto Y_t e^{\theta t} \end{aligned} \tag{3.29}$$

To look for simplifications, we thus define another random variable Z_t as a function of Y_t :

$$Z_t = f\left(t, \theta, Y_t\right) = e^{\theta t} Y_t \tag{3.30}$$

Using Ito's Lemma in equation (3.25), it follows:

$$df\left(t, Y_t\right) = \left(\frac{\partial f}{\partial Y_T}\right) dY_t + \left(\frac{\partial f}{\partial t}\right) dt + \frac{1}{2}\left(\frac{d^2 f}{dY_T^2}\right) \left(dY_T\right)^2 \tag{3.31}$$

Where note that $\frac{d^2 f}{dY_T^2} = \frac{d^2}{dY_T^2}\left[e^{\theta t} Y_t\right] = 0$. Thus, Ito formula assumes the following form in our case:

$$df\left(\theta, Y_t\right) = \left(\frac{\partial f}{\partial Y_T}\right) dY_t + \left(\frac{\partial f}{\partial t}\right) dt \tag{3.32}$$

Applying this,

$$dZ\left(\theta, Y_t\right) = \theta e^{\theta t} Y_t dt + e^{\theta t} dY_t$$

Plugging equation (3.28),

$$\begin{aligned} dZ(\theta, Y_t) &= \theta e^{\theta t} Y_t dt + e^{\theta t}\left(-\theta Y_t dt + \sigma dW_t\right) \\ &= e^{\theta t} \sigma dW_t \end{aligned} \tag{3.33}$$

Equation (3.33) is easy to solve. To do so we merely sum all the stochastic increments dW over a finite time t , noting that each one is multiplied by $e^{\theta t}\sigma$. Thus, the integral form is.

$$\begin{aligned} Z_t &= Z_s + \int_s^T dZ_t \\ &= Z_s + \sigma \int_s^T e^{\theta t} dW_t \end{aligned} \tag{3.34}$$

Where S is the start of the integration through time. Now that we found a solution to the random variable Z_t it is time to go back through the substitutions to find the solution to X_t. In order to achieve that we first reverse the exponential component in the relationship between Y_t and Z_t :

$$Y_t = e^{-\theta t} Z_t$$

$$Y_T = e^{-\theta T}\left(Z_S + \sigma \int_S^T e^{\theta t} dW_t\right)$$
$$Y_T = e^{-\theta T}\left(e^{kS} Y_S + \sigma \int_S^T e^{\theta t} dW_t\right)$$
$$Y_T = e^{-\theta(T-S)} Y_S + \sigma \int_S^T e^{\theta(t-T)} dW_t. \tag{3.35}$$

Finally plugging $Y_t = X_t - \mu$:

$$X_T - \mu = e^{-\theta(T-S)}\left(X_S - \mu\right) + \sigma \int_S^T e^{\theta(t-T)} dW_t$$
$$X_T = \mu + e^{-\theta(T-S)}\left(X_S - \mu\right) + \sigma \int_S^T e^{\theta(t-T)} dW_t$$

Starting from $S = 0$, we obtain:

$$X_T = \mu + e^{-\theta T}\left(X_0 - \mu\right) + \sigma \int_{s=0}^T e^{-\theta(T-s)} dW_s \tag{3.36}$$

To completely determine X_T in (3.36), note that the stochastic integral represents the sum of Gaussian random variables. Thus all we need to do is calculate its mean and variance:

Mean of $X(t)$ in (3.36)

$$\mathbb{E}\left[X_t\right] = \mathbb{E}\left[\mu + e^{-\theta T}\left(X_0 - \mu\right) + \sigma \int_{s=0}^T e^{-\theta(T-t)} dW_s\right]$$
$$\mathbb{E}\left[X_t\right] = \mu + e^{-\theta T}\left(X_0 - \mu\right) + \mathbb{E}\left[\sigma \int_{s=0}^T e^{-\theta(T-t)} dW_s\right]$$

Where $\int_{S=0}^T e^{-\theta(T-t)} dW_s$ is a Weiner process. Thus, $\mathbb{E}\left[\sigma \int_{S=0}^T e^{-\theta(T-t)} dW_S\right] = 0$.

Therefore, the mean of the stochastic integral in (3.36) is the following:

$$\mathbb{E}\left[X_t\right] = \mu + e^{-\theta T}\left(X_0 - \mu\right) \tag{3.37}$$

Variance of $X(t)$ in (3.36)

$$\mathbb{V}(Z_t) = \mathbb{E}\left[(X_t - \mathbb{E}[X_t])^2\right]$$
$$\mathbb{V}(Z_t) = \mathbb{E}\left[\left(\mu + e^{-\theta T}(X_0 - \mu) + \sigma \int_{s=0}^{T} e^{-\theta(T-s)} dW_s - \mathbb{E}[X_t]\right)^2\right]$$

Plugging in (3.37) inside the expression,

$$\mathbb{V}(X_t) = \mathbb{E}\left[\left(\sigma \int_{s=0}^{T} e^{-\theta(T-s)} dW_s\right)^2\right]$$

From Ito isometry, $dW^2 = dt$,

$$\mathbb{V}(X_t) = \sigma^2 \int_{s=0}^{T} e^{-2\theta(T-s)} ds$$
$$\mathbb{V}(X_t) = \sigma^2 \left[\frac{1}{2\theta} e^{-2\theta(T-s)}\right]_{s=0}^{t}$$
$$\mathbb{V}(X_t) = \frac{\sigma^2}{2\theta}\left(1 - e^{-2\theta t}\right). \tag{3.38}$$

Applying the limit $t \to \infty$ for (3.37) and (3.38) allows us to recover the stationary distribution:

$$\lim_{t\to 0} \mathbb{E}[X_t] = \lim_{t\to\infty} \mu + e^{-\theta t}(X_0 - \mu) = \mu \tag{3.39}$$

$$\lim_{t\to 0} \mathbb{V}[X_t] = \lim_{t\to\infty} \frac{\sigma^2}{2\theta}\left(1 - e^{-2\theta t}\right) = \frac{\sigma^2}{2\theta} \tag{3.40}$$

3.4 The Full Linear Stochastic Equation

Geometric Brownian Motion

We now extend our analysis from (3.26) to solving the more general stochastic linear equation:

$$dS_t = -\mu_t S_t \, dt + \sigma_t S_t \, dW_t \tag{3.41}$$

where $S_t(W_t)$ do not have explicit time dependence. This equation is called the **Geometric Brownian Motion**. We can write the above as following:

$$\frac{dS_t}{S_t} = \mu_t \, dt + \sigma_t \, dW_t \tag{3.42}$$

Notice that as S_t approaches zero, so does the change (else the LHS blows up). This effectively limits S_t to positive values, $S_t > 0$.

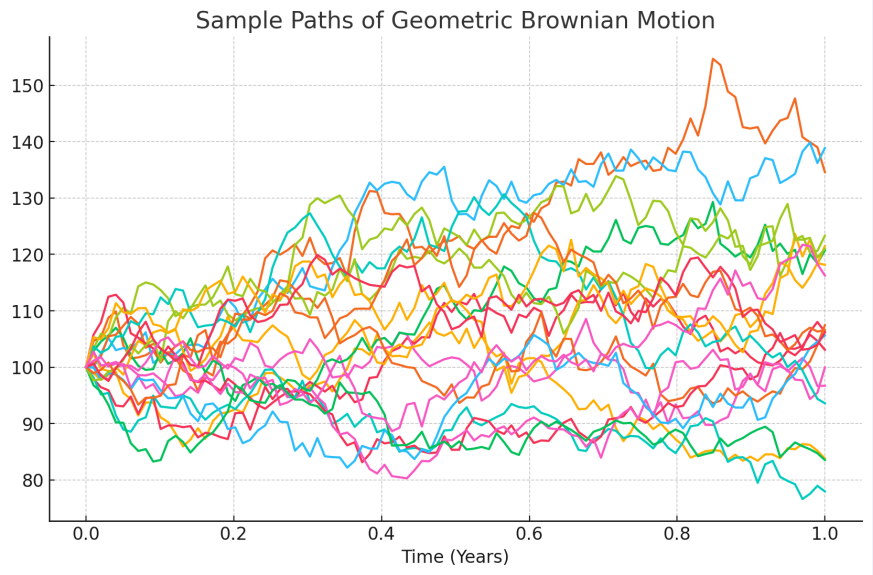


Figure 2: Sample Paths obeyings Geometric Brownian Motion

The reason we solve this equation is because it introduces us to a **stochastic version of the log-derivative trick**, a technique which would come handy when we go over Anderson's rule.

To make headways into solving 3.42, we notice that the quantity $\frac{dS_t}{S_t}$ has striking similarity to:

$$\frac{\partial \ln S(x)}{\partial x} = \frac{1}{S(x)} \frac{\partial S(x)}{\partial x} \tag{3.43}$$

But since we are working with stochastic processes, we can't apply regular calculus.

To come up with a stochastic version of the log-derivative, we utilize Ito's Lemma as stated in (3.25):

$$d\ln S_t = \frac{\partial \ln S_t}{\partial t}dt + \frac{\partial \ln S_t}{\partial S_t}dS_t + \frac{1}{2}\frac{\partial^2 \ln S_t}{\partial S_t^2}dS_t^2 \tag{3.44}$$

The derivatives are:

$$\frac{\partial S_t(W_t)}{\partial t} = 0, \quad \frac{\partial \ln S_t}{\partial S_t} = \frac{1}{S_t}, \quad \frac{\partial^2 \ln S_t}{\partial S_t^2} = -\frac{1}{S_t^2}.$$

Equation (3.44) can then be written as:

$$d\ln S_t = \frac{1}{S_t}dS_t - \frac{1}{2}\frac{1}{S_t^2}dS_t^2. \tag{3.45}$$

Plugging in equation (3.41),

$$\begin{aligned}
d\ln S_t &= \frac{1}{S_t}dS_t - \frac{1}{2}\frac{1}{S_t^2}\left(-\mu_t S_t\,dt + \sigma_t S_t\,dW_t\right)^2 \\
d\ln S_t &= \frac{1}{S_t}dS_t - \frac{1}{2}\frac{1}{S_t^2}(\mu_t^2 S_t^2\,dt^2 + \sigma_t^2 S_t^2\,dW_t^2 - 2\mu_t S_t \sigma_t S_t\,dW_t\,dt) \\
d\ln S_t &= \frac{1}{S_t}dS_t - \frac{1}{2}\frac{1}{S_t^2}(\sigma_t^2 S_t^2\,dW^2)
\end{aligned} \tag{3.46}$$

Where we ignored the cross-differentials and dt^2 in the last step. We then have using Ito's rule:

$$\begin{aligned}
d\ln S_t &= \frac{1}{S_t}dS_t - \frac{1}{2}\sigma_t^2\,dt \\
\frac{1}{S_t}dS_t &= d\ln S_t + \frac{1}{2}\sigma_t^2\,dt.
\end{aligned} \tag{3.47}$$

Substituting this into equation (3.42):

$$\begin{aligned}
\mu_t dt + \sigma_t dW_t &= d\ln S_t + \frac{1}{2}\sigma_t^2 dt \\
\int_0^t d\ln S_t &= \int_0^t \mu_s\,ds + \int_0^t \frac{1}{2}\sigma_s^2\,ds + \int_0^t \sigma_s\,dW_s \\
\ln S_t - \ln S_0 &= \mu_t t - \frac{1}{2}\sigma_t^2 t + \sigma_t W_t \\
\ln \frac{S_t}{S_0} &= \mu_t t - \frac{1}{2}\sigma_t^2 t + \sigma_t W_t \\
S_t &= S_0 \exp\left[\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W_t\right]
\end{aligned} \tag{3.48}$$

Where (3.48) represents our final solution.

3.5 Auto-correlation of Stochastic Processes

Stationary Stochastic Process

A stochastic process $X(t)$ is called **stationary** if its statistical properties do not change over time. In a strict sense, this means that the joint distribution of $X(t_1), X(t_2), \ldots, X(t_n)$ depends only on the relative time differences (lags) between $t_1, t_2, \ldots, t_n$, and not on the absolute times themselves.

Autocorrelation Function and Correlation Coefficient

The autocorrelation function (ACF) of a stationary process is defined as:

$$\rho(\tau) = \frac{\text{Cov}(X(t), X(t+\tau))}{\sigma^2}, \tag{3.49}$$

The ACF $\rho(\tau)$ measures the strength of the linear relationship between $X(t)$ and $X(t+\tau)$. It is normalized to have values in the range $[-1, 1]$, where 1 indicates perfect positive correlation at lag τ, -1 indicates perfect negative correlation at lag τ and 0 indicates no linear correlation at lag τ.

The **correlation coefficient** quantifies the strength and direction of a linear relationship between two variables. It is a normalized value that ranges between -1 and 1 with $+1$ indicating perfect correlation, -1 indicating negative correlation and 0 indicating no correlation. The formula assumes the form:

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}, \tag{3.50}$$

where $\text{Cov}(X, Y)$ is the covariance between X and Y, and σ_X and σ_Y are the standard deviations of X and Y respectively

Let us now calculate the correlation coefficient for the Wiener Process. We know already that $\mathbb{V}(W(t)) = t$ and thus $\mathbb{V}(W(t+\tau)) = \text{t} + \tau$. We can then calculate the correlation $\langle W(t)W(t+\tau)\rangle$ in the following way:

$$\langle W(t)W(t+\tau)\rangle = \left\langle \int_0^t dW \int_0^{t+\tau} dW \right\rangle = \left\langle \int_0^t dW \left(\int_0^t dW + \int_t^{t+\tau} dW \right) \right\rangle$$

$$\langle W(t)W(t+\tau)\rangle = \left\langle \left(\int_0^t dW\right)^2 + \int_0^t dW \int_t^{t+\tau} dW \right\rangle$$
$$\langle W(t)W(t+\tau)\rangle = \left\langle \left(\int_0^t dW\right)^2 \right\rangle + \left\langle \int_0^t dW \int_t^{t+\tau} dW \right\rangle$$
$$\langle W(t)W(t+\tau)\rangle = \left\langle (W(t))^2 \right\rangle + \left\langle \int_0^t dW \right\rangle \left\langle \int_t^{t+\tau} dW \right\rangle$$
$$\langle W(t)W(t+\tau)\rangle = \left\langle (W(t))^2 \right\rangle + \langle W(t)\rangle\langle W(\tau)\rangle$$
$$\langle W(t)W(t+\tau)\rangle = t + 0 = t. \tag{3.51}$$

Thus, the correlation coefficient is given as following:

$$C_{X(t)X(t+\tau)} = \frac{t}{\sqrt{t(t+\tau)}} = \sqrt{\frac{1}{(1+\tau/t)}} \tag{3.52}$$

Where we used the fact that the random variables $\mathrm{A} = \int_0^t dW$ and $\mathrm{B} = \int_t^{t+\tau} dW$ are independent, which implies their correlation $\langle AB\rangle$ is just the product of their means, $\langle A\rangle\langle B\rangle$. As expected, the Wiener process at time $\mathrm{t}+\tau$ is increasingly independent of its value at an earlier time t as τ increases.

Two-time autocorrelation Function

The two-time correlation function is defined as:

$$\mathrm{g}\,(\mathrm{t},\mathrm{t}') = \langle \mathrm{X(t)X}\,(\mathrm{t}')\rangle \tag{3.53}$$

If the mean of the process X(t) is constant with time, and the two-time auto-correlation function, $\mathrm{g}(\mathrm{t},\mathrm{t}+\tau) = \langle \mathrm{X(t)X(t}+\tau)\rangle$ is also independent of the time, t , so that it depends only on the timedifference, τ, then X(t) is referred to as being "wide-sense" stationary. In this case, the two-time auto-correlation function depends only on τ, and we write:

$$g(\tau) = \langle \mathrm{X(t)X}\,(\mathrm{t}')\rangle \tag{3.54}$$

The auto-correlation function for a wide-sense stationary process is always symmetric, so that $\mathrm{g}(-\tau) = \mathrm{g}(\tau)$. This is easily shown by noting that:

$$g(\tau) = \langle \mathrm{X(t)X(t}-\tau)\rangle = \langle \mathrm{X(t)X(t}+\tau)\rangle = g(\tau) \tag{3.55}$$

One can calculate the two-time autocorrelation for some arbitrary process $X(t)$ as long has one has joint probability density $P(x',t';x,t)$. Let us define the probability density as,

$$P\left(x,t;x',t'\right)=P\left(x,t\mid x',t'\right)P\left(x',t'\right) \tag{3.56}$$

The conditional probability is the probability density for X at time t , given that X has the value x ' at time t '. In fact, we already know how to calculate this, since it is the same thing that we have been calculating all along in solving stochastic differential equations: the solution to an SDE for X is the probability density for X at time t , given that its initial value at t = 0 is x_0. To obtain the conditional probability in 3.56, all we need to do is solve the SDE for x but this time with the initial time being t' rather than 0.

As an example, let us do this for the simplest stochastic equation, d$\mathbf{X}$ = d$\mathbf{W}$. Solving the SDE means summing all the increments dW from time t$'$ to t , with the initial condition $X\left(t'\right)=x'$. The solution is,

$$X(t)=x'+\int_{t'}^{t}dW=x'+W\left(t-t'\right) \tag{3.57}$$

And this has the probability density,

$$P\left(x,t\mid x',t'\right)=P(x,t)=\frac{e^{-(x-x')^2/[2(t-t')]}}{\sqrt{2\pi\left(t-t'\right)}} \tag{3.58}$$

To calculate the joint probability density we now need to specify the density for X at time t$'$. If X started with the value 0 at time 0 , then at time t' the density for X (t$'$) is just the density for the Wiener process, thus,

$$P\left(x',t'\right)=\frac{e^{-\frac{x'^2}{[2t^2]}}}{\sqrt{2\pi}t'} \tag{3.59}$$

Using equation (3.58) and (3.59), the joint probability density is,

$$P\left(w,t\mid w',t'\right)=\frac{e^{-\frac{(x-x')^2}{[2(t-t')]-\frac{x'^2}{[2t']}}}}{\sqrt{2\pi\left(t-t'\right)t'}} \tag{3.60}$$

And the correlation function is therefore,

$$\left\langle X\left(t'\right)X(t)\right\rangle=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xx'P\left(x,t;x',t'\right)dxdx'=t' \tag{3.61}$$

We then obtain the correlation coefficient between X(t) and X (t$'$) by dividing this by the square root of the product of the variances as above

4 Properties of Stochastic Processes

4.1 Fourier Domain of Stochastic Processes

Definition of Fourier Transform of Stochastic Processes

Consider a stochastic process $\mathbf{x}(t)$, which has many possible sample paths. This process can be described by a probability density function over the collection of possible sample paths. Each sample path is a function of time, denoted as $\mathbf{x}_\alpha(t)$, where α labels the different possible paths. Thus, $\mathbf{x}(t)$ is a random function, whose possible values are the functions $\mathbf{x}_\alpha(t)$ (see figure below)

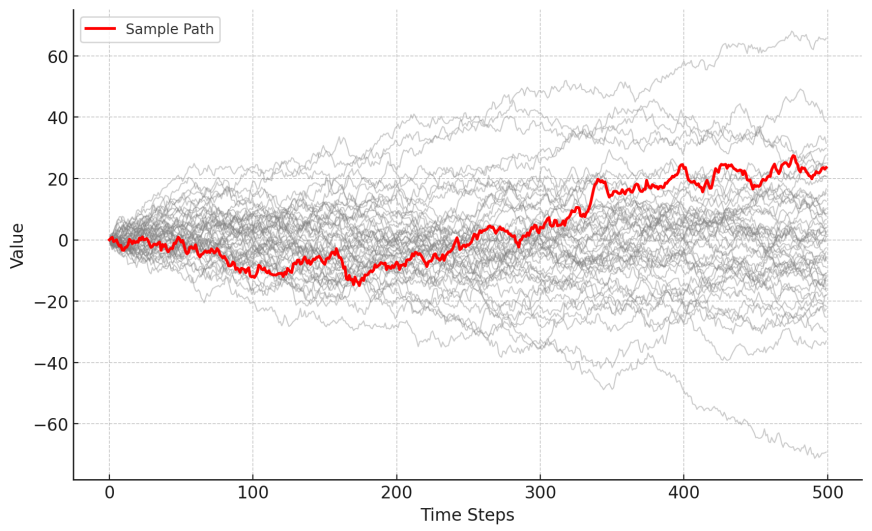


Figure 3: Realizations of a Stochastic Process with a Highlighted Sample Path

Similarly, the Fourier transform $\mathbf{X}(\omega)$ of the stochastic process is defined as a random function, whose possible values are the Fourier transforms of the individual sample paths. The possible values of $\mathbf{X}(\omega)$ are given by:

$$\mathbf{X}_\alpha(\omega) = \int_{-\infty}^{\infty} \mathbf{x}_\alpha(t) e^{-i2\pi\omega t}\, dt. \tag{4.1}$$

Energy Conservation in Stochastic Signals

For a stochastic signal, $\mathbf{x}(t)$, the total average energy in the signal is the average value of the instantaneous power, $\mathbf{x}^2(t)$, integrated over all time:

$$\mathrm{S}[\mathbf{x}(t)] = \int_{-\infty}^{\infty} \left\langle \mathbf{x}(t)^2 \right\rangle \, \mathrm{d}t. \tag{4.2}$$

With the total energy being conserved in the ω-space:

$$\mathrm{S}[\mathbf{X}(\omega)] = \int_{-\infty}^{\infty} \left\langle |\mathbf{X}(\omega)|^2 \right\rangle \, \mathrm{d}\omega. \tag{4.3}$$

Claim: The autocorrelation and the energy spectrum of the stochastic process are Fourier pairs

For a deterministic signal $f(t)$, it is well-known that the autocorrelation function $g(\tau)$ and the energy spectrum $S(\omega)$ are related as a Fourier transform pair. With the definition of the Fourier transform of a stochastic process, we can now derive the proof that the autocorrelation and the energy spectrum of the stochastic process are Fourier pairs. Let us define the autocorrelation function as:

$$g(t, \tau) = \langle x(t)x(t+\tau) \rangle \tag{4.4}$$

Then,

$$\begin{aligned}
\langle |X(v)|^2 \rangle &= \left\langle \int_{-\infty}^{\infty} x(t) e^{-i2\pi\omega t} dt \int_{-\infty}^{\infty} x(t) e^{i2\pi\omega t} dt \right\rangle \\
\langle |X(v)|^2 \rangle &= \left\langle \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x(t)x(t-\tau)dt) e^{-i2\pi\omega t} d\tau \right\rangle \\
\langle |X(v)|^2 \rangle &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \langle x(t)x(t-\tau)dt \rangle e^{-i2\pi\omega t} d\tau \\
\langle |X(v)|^2 \rangle &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(t, \tau) e^{-i2\pi\omega t} d\tau
\end{aligned} \tag{4.5}$$

Average Power & Power Spectral Density of a Stochastic Signal

Average Power: We define average power of a stochastic signal in the same way as for a deterministic signal. However this time we take the expectation value of the process so as to average the power both over time and over all realizations (sample paths) of the process. Thus the average power of a stochastic process is:

$$S[x(t)] = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} \left\langle x^2(t) \right\rangle dt \tag{4.6}$$

Power spectral density: This is defined as the power per unit frequency of a sample path of the process, averaged over all the sample paths:

$$S[x(t)] = \int_{v_1}^{v_2} S(v) dv \tag{4.7}$$

Wiener-Khinchin theorem

The power spectral density of a wide-sense stationary stochastic process $x(t)$ is the Fourier transform of the two-time auto-correlation function:

$$S(v) = \int_{-\infty}^{\infty} g(\tau) e^{-2\pi v\tau} d\tau, g(\tau) = \langle x(t)x(t+\tau) \rangle \tag{4.8}$$

4.2 White Noise

The power spectra in Fourier domain of noise signals are described using the concept of **colors**. For example, **white noise** refers to a noise signal with a spectral density that is uniform across all frequencies, also known as a flat power spectral density. The name *white* is derived from white light, as it contains all colors in the visible spectrum. Each component of white noise has a probability distribution with zero mean, finite variance, and statistical independence. On the other hand, **red noise** arises from Brownian motion. Its spectral density is inversely proportional to the square of the frequency, meaning that its power decreases significantly as frequency increases. This leads to more energy at low frequencies.

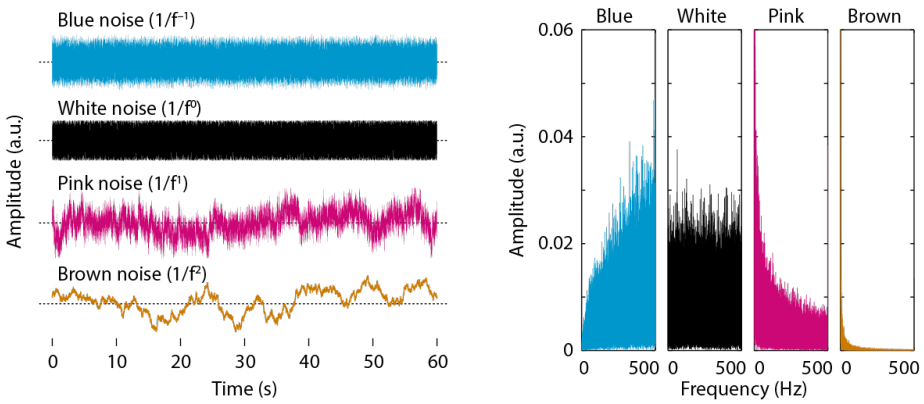


Figure 4: Different realizations of noise

Let us try to understand white noise more concretely. Consider a function $f(t)$ whose integral from $-\infty$ to ∞ is finite in position space. The more sharply peaked f (that is, the smaller the smallest time interval containing the majority of its energy), then the less sharply peaked is its Fourier transform. Similarly, the broader a function, then the narrower is its Fourier transform. Now consider a stochastic process $x(t)$. If the auto-correlation function $g(\tau) = \mathrm{x(t)x(t+\tau)}$ drops to zero very quickly as $|\tau|$ increases, then the power spectrum of the process x(t) must be broad, meaning that x(t) contains high frequencies. This is reasonable, since if a process has high frequencies it can vary on short time-scales, and therefore become uncorrelated with itself in a short time.

We know that the paths of $W(t)$ are too irregular to have a well-defined derivative. However, in the stochastic calculus framework, we can define an idealized derivative called a white noise process $\eta(t)$:

$$\eta(t) = \frac{dW}{dt} \tag{4.9}$$

This can be useful as a calculational tool. Since the increments of the Wiener process in two consecutive time intervals dt are independent of each other, η(t) must be uncorrelated with itself whenever the time separation is greater than zero. Thus we must have $\langle \eta(t)\eta(t+\tau)\rangle = 0$ if $\tau > 0$. In addition, if we try to calculate $\langle \eta(t)\eta(t+\tau)\rangle$, we obtain:

$$\begin{aligned} g(0) = \langle \eta(t)\eta(t+\tau)\rangle &= \lim_{\Delta t\to 0}\left\langle \frac{\Delta W}{\Delta t}\frac{\Delta W}{\Delta t}\right\rangle \\ &= \lim_{\Delta t\to 0}\left\langle \frac{(\Delta W)^2}{(\Delta t)^2}\right\rangle = \lim_{\Delta t\to 0}\frac{1}{\Delta t} = \infty \end{aligned} \tag{4.10}$$

A function which has this property is the delta-function $\delta(\tau)$. Let us see what happens if we assume that η(t) is a noise source with the autocorrelation function:

$$\frac{dW}{dt} = \langle \eta(t)\eta(t+\tau)\rangle = \delta(\tau) \tag{4.11}$$

We now demonstrate how we can use idealized noise to solve stochastic differential equations and compute quantities like variance and correlation. In particular, we solve the equation:

$$dx = \sigma dW \tag{4.12}$$

The solution to (4.12) is x(t) = σW(t), as we already worked out. If η(t) exists then we can write the stochastic equation as $dW = \eta(t)dt$:

$$\begin{aligned} dx &= \sigma\eta(t)dt \\ \frac{dx}{dt} &= \sigma\eta(t) \\ x(t) &= \sigma \int_0^t \eta(s)ds \end{aligned} \tag{4.13}$$

Let us calculate the variance of x(t). This is:

$$\mathbb{V}(x(t)) = \left\langle x(t)^2 \right\rangle - \langle x(t) \rangle^2$$

Where $\langle x(t) \rangle = 0$ as η(t) is a zero-mean process. Then, from equation (4.13)

$$\begin{aligned} V(x(t)) &= \left\langle \sigma^2 \int_0^t \delta(s)ds \int_0^t \delta(v)dv \right\rangle \\ &= \sigma^2 \int_0^t \int_0^t \langle \delta(s)\delta(v) \rangle dsdv \\ &= \sigma^2 \int_0^t \int_0^t \delta(s-v)dsdv = \sigma^2 \int_0^t dv = \sigma^2 t \end{aligned} \tag{4.14}$$

which is the correct answer. Also, we can calculate the two-time auto-correlation function of $x(t)$ using (4.11):

$$\begin{aligned} \langle x(t)x(t+\tau) \rangle &= \left\langle \sigma^2 \int_0^t \delta(s)ds \int_0^{t+\tau} \delta(v)dv \right\rangle = \sigma^2 \int_0^t \int_0^{t+\tau} \langle \delta(s-v) \rangle dsdv \\ &= \sigma^2 \int_0^t dv = \sigma^2 t \end{aligned} \tag{4.15}$$

which is also correct. To sum up, we have been able to obtain the correct solution to the stochastic differential equation by assuming that:

$$\eta(t) \equiv \frac{dW(t)}{dt}$$

exists and has a delta auto-correlation function.

This technique will work for any SDE that has purely additive noise, but it does not work for SDEs in which the noise dW multiplies a function of any of the variables. For example, one cannot use it to solve the equation $\mathbf{dx} = \mathbf{xdW}$.

Proof that Power Spectrum of White Noise is constant

We can use (4.11) to obtain the characteristic nature of white noise in the power spectrum. The Fourier Transform of $\eta(t)$ is given by:

$$\tilde{\eta}(f) = \int_{-\infty}^{\infty} \eta(t) e^{-i2\pi ft}\, dt. \tag{4.16}$$

The power spectrum is defined as the squared magnitude of the Fourier Transform and averaged over realizations:

$$S_\eta(f) = \langle |\tilde{\eta}(f)|^2 \rangle. \tag{4.17}$$

Substituting the Fourier Transform of $\eta(t)$:

$$|\tilde{\eta}(f)|^2 = \left| \int_{-\infty}^{\infty} \eta(t) e^{-i2\pi ft}\, dt \right|^2 \tag{4.18}$$

Taking the ensemble average:

$$S_\eta(f) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \langle \eta(t)\eta(t') \rangle e^{-i2\pi f(t-t')}\, dt\, dt'. \tag{4.19}$$

Using the delta correlation property in (4.11), the integral simplifies:

$$S_\eta(f) = \int_{-\infty}^{\infty} 2D\delta(t-t') e^{-i2\pi f(t-t')}\, dt\, dt'. \tag{4.20}$$

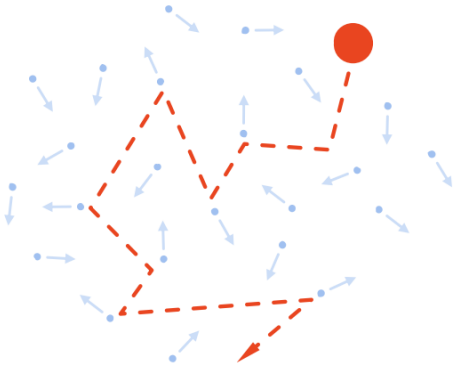
The Dirac delta function selects $t = t'$, reducing the integral:

$$S_\eta(f) = 2D \int_{-\infty}^{\infty} e^{-i2\pi f(t-t)}\, dt = 2D. \tag{4.21}$$

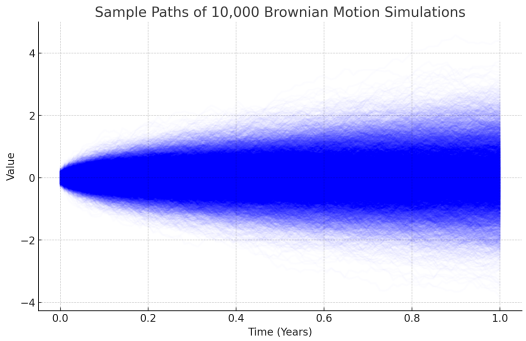
Where D is a constant. Thus, the power spectrum of white noise is:

$$S_\eta(f) = \text{constant}. \tag{4.22}$$

5 Brownian Motion:



(a) Physical Illustration of Motion



(b) 10,000 Brownian Motion Paths

5.1 Introduction to Brownian Motion

Classically, it is difficult to explain the haphazard manner in which the pollen grains inside the fluid behave. In Newtonian Mechanics, force is required to alter the state of motion. However, in the case of pollen grains, it was not as obvious as to what exactly is the cause of the 'force' that led these grains to move in a haphazard manner. Before Einstein's paper in 1905, the plausible explanation was thought to be they might possess some "vital force" or were perhaps biologically active. This hypothesis implied that the motion was self-generated rather than influenced by external factors.

However, Albert Einstein demonstrates that no active mechanism is necessary and that the random forces generated by the thermally excited water molecules can account for the motion of the grains. This explanation was confirmed by Jean Perrin in 1908, for which he was awarded the Nobel prize in 1926. We will now give a gist of the argument that explains the shift that Einstein introduced. Consider the particle's position represented by a single coordinate x. For a particle submerged in a fluid, the forces on it include external potential forces (such as gravity) and a resistive frictional force arising from the fluid's viscosity. The deterministic equation describing its motion is:

$$m\ddot{x} = -\frac{\partial V}{\partial x} - \frac{1}{\mu}\dot{x} \tag{5.1}$$

However, in many cases, especially for small particles moving in a viscous medium, the inertial term ($m\ddot{x}$) becomes negligible compared to the viscous and external potential forces. To understand why, we analyze the relative importance of the inertial and viscous terms using the **Reynolds number (Re)**, a dimensionless parameter that compares inertial forces to viscous forces:

$$Re = \frac{\rho a v}{\eta} \tag{5.2}$$

where ρ is the particle's density, a is its characteristic size,v is its velocity and η is the fluid's viscosity. Inertial forces (F_{inertial}) scale as:

$$F_{\text{inertial}} \sim \frac{\rho a^2 v^2}{a} = \rho a v^2 \tag{5.3}$$

On the other hand, viscous forces (F_{viscous}) scale as:

$$F_{\text{viscous}} \sim \eta a v \tag{5.4}$$

For small particles, particularly at microscopic scales, the Reynolds number is extremely small ($Re \ll 1$), indicating that viscous forces dominate over inertial forces. While the Reynolds number of macroscopic objects like swimming animals is high and therefore inertia cannot be ignored, a bacterium (approximately $1\,\mu$m) swimming at $30\,\mu$m/s has $Re \approx 10^{-4}$. At low Reynolds numbers, typical for subcellular and molecular systems, the inertial term $m\ddot{x}$ in the equation of motion can be neglected. This simplifies the governing equation:

$$-\frac{\partial V}{\partial x} - \frac{1}{\mu}\dot{x} = 0 \tag{5.5}$$

Thus, the velocity of the particle ($\dot{x}$) is directly proportional to the external force acting on it because the system is dominated by viscous interactions rather than inertia.

Einstein's insight in the 1905 paper was to realize the **particular nature of water**. At macroscopic scales, we often treat fluids like water as continuous mediums governed by classical fluid dynamics. However, as we zoom into microscopic scales (microns and below), this approximation breaks down. At these scales the particulate nature of water becomes apparent and the interactions between water molecules and the particle are no longer smooth but instead involve collisions. Let us denote the random force exerted on the particle by these water molecules as $\eta(t)$. Then, (5.5) assumes the form:

$$\dot{x} = -\mu\frac{\partial V}{\partial x} + \eta(t) \tag{5.6}$$

The stochastic (5.7) is the Langevin equation for the coordinate x in the **underdamped regime**. For the **overdamped regime**, the inertial term $m\ddot{x}$ would not be negligible such that the equation would read:

$$m\ddot{x} = -\dot{x} - \mu\frac{\partial V}{\partial x} + \eta(t) \tag{5.7}$$

Where different realizations of the stochastic force $\eta(t)$ lead to different values of x(t).

Langevin Equation in various regimes

The general **damped** Langevin equation reads:

$$m\frac{d^2x}{dt^2} + \gamma\frac{dx}{dt} + \frac{dV(x)}{dx} = \eta(t), \tag{5.8}$$

In the **overdamped regime**, inertia is negligible ($m \approx 0$), so the equation simplifies to:

$$\gamma\frac{dx}{dt} + \frac{dV(x)}{dx} = \eta(t), \tag{5.9}$$

In the **critically damped regime**, the damping coefficient is tuned to the critical value:

$$\gamma_{\text{crit}} = 2\sqrt{mk}. \tag{5.10}$$

For a general potential $V(x)$, this leads to:

$$m\frac{d^2x}{dt^2} + 2\sqrt{m\frac{d^2V(x)}{dx^2}}\frac{dx}{dt} + \frac{dV(x)}{dx} = \eta(t). \tag{5.11}$$

Constraints on Random Force $\eta(t)$

1. The net force from the collision of a submerged particle with these water molecules should average to zero over time because of the uniform distribution of impacts. As a result, the **mean** of $\eta(t)$ should be 0:

$$\langle \eta(t) \rangle = 0 \tag{5.12}$$

2. The impacts at one time are uncorrelated with impacts at another time, provided the times are sufficiently separated. This can be expressed in terms of the two-time auto-correlation function of random force $\eta(t)$:

$$\langle \eta(t)\eta\left(t'\right) \rangle = 2D\delta\left(t - t'\right) \tag{5.13}$$

   Where D is the measure of the measure of the strength of the fluctuating force. The diract delta function ensures that correlation occurs only when $t = t'$ and otherwise it is 0.

5.2 Solving Brownian Equation

From fluid dynamics, a particle such as a pollen grain experiences a frictional force when moving through a viscous fluid. This force is proportional to the negative of the particle's momentum:

$$F_{\text{friction}} = -\gamma \text{p} = -\gamma m v \tag{5.14}$$

Where m is the mass of the grain and v is its velocity. The constant of proportionality, γ, is usually referred to as the damping rate, and is given by $\gamma = 6\pi\frac{na}{m}$. Restricing ourselves to one-dimensional case, Newton's 2nd Law becomes:

$$m\frac{d^2x}{dt} = F_{\text{friction}} + F_{\text{fluct}}$$

$$m\frac{d^2x}{dt} = -\gamma p + F_{\text{fluct}} \tag{5.15}$$

From our discussion in the previous section, we introduce a random force $\eta(t)$ characterized by the following autocorrelation function:

$$\langle \eta(t)\eta(t+\tau)\rangle = \delta(\tau)$$

Thus, the equation above becomes:

$$\frac{d^2x}{dt^2} = -\gamma p + \sigma\eta(t) \tag{5.16}$$

Where σ controls the strength of fluctuation. Note that our discussion for now has not assumed that **intertial term is negligible**. Using p = m $\frac{dx}{dt}$,

$$\frac{dx}{dt} = \frac{p}{m}, \tag{5.17}$$

$$\frac{d^2x}{dt^2} = \frac{dp}{dt} = -\gamma p + \sigma\eta(t). \tag{5.18}$$

In vector form:

$$\frac{d}{dt}\begin{pmatrix} x \\ p \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{m} \\ 0 & -\gamma \end{pmatrix}\begin{pmatrix} x \\ p \end{pmatrix} + \begin{pmatrix} 0 \\ \sigma\eta(t) \end{pmatrix} \tag{5.19}$$

Assume now that the thermal noise is randomly distributed s.t $\frac{dW}{dt} = \eta(t)$:

$$\frac{d}{dt}\begin{pmatrix} x \\ p \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{m} \\ 0 & -\gamma \end{pmatrix}\begin{pmatrix} x \\ p \end{pmatrix} + \begin{pmatrix} 0 \\ \sigma \end{pmatrix}\frac{dW}{dt}$$

$$\begin{pmatrix} dx \\ dp \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{m} \\ 0 & -\gamma \end{pmatrix}\begin{pmatrix} x \\ p \end{pmatrix}dt + \begin{pmatrix} 0 \\ \sigma \end{pmatrix}dW \tag{5.20}$$

To obtain the solution for the system of equations above, we note that dx depends upon p but dp does not depend on x . Therefore, we first seek to solve p

Solution for $p(t)$

Let us first solve for the equation:

$$dp = -\gamma p dt + \sigma dW \tag{5.21}$$

This equation represents Ornstein-Uhlenbeck process as characterized by equation (3.26). We already solved this system to find the expectation and variance. The solution reads:

$$p(t) = e^{-\theta(t-S)} p_s + \sigma \int_S^t e^{-\gamma(t-T)} dW_t$$

Where for $S = 0$:

$$p(t) = e^{-\gamma t} p(0) + \sigma \int_0^t e^{-\gamma(t-T)} dW_t \tag{5.22}$$

The Expectation and variance become:

$$\begin{aligned} \mathbb{E}\left[p_t\right] &= e^{-\gamma t} p_s, \\ \mathbb{V}\left[p_t\right] &= \frac{\sigma^2}{2\gamma}\left(1 - e^{-2\gamma t}\right). \end{aligned} \tag{5.23}$$

The Equipartition theorem states that the average kinetic energy of the particle is related to the temperature

$$\langle E \rangle = \left\langle p^2/2m \right\rangle = \frac{k_B T}{2} \tag{5.24}$$

For this to be satisfied, we require that $p(t)$ in equation (5.23) be independent of time as $t \to \infty$:

$$\lim_{t\to\infty} \mathbb{V}(p(t)) = \frac{\sigma^2}{2\gamma} = \mathbb{V}(p(t))_{SS} \tag{5.25}$$

Where the subscript SS indicates steady state. Since the mean value of $p(t)$ tends to zero as $t \to \infty$, in the steady-state the mean is zero. This leads to the simplification:

$$\begin{aligned} \mathbb{V}(p(t)) &= \left\langle p^2 \right\rangle - \langle p \rangle^2 \\ \mathbb{V}(p(t)) &= \left\langle p^2 \right\rangle \end{aligned} \tag{5.26}$$

For consistently with statistical mechanics, we require,

$$\langle E \rangle = \left\langle p^2/2m \right\rangle = \frac{\mathbb{V}(p)}{2m} = \frac{\sigma^2}{4\gamma m} = \frac{\sigma^2}{24\pi\eta d} \tag{5.27}$$

Equating this with $k_B T/2$, we see that the strength of the noise must be,

$$\sigma = \sqrt{12\pi\eta k T} \tag{5.28}$$

Solution for $x(t)$

We now turn to the second equation x in (5.20):

$$x(t) = \frac{1}{m}\int_0^t p(s)ds \tag{5.29}$$

From equation (5.22),

$$\begin{aligned}
x(t) &= \frac{1}{m}\int_0^t \left[e^{-\gamma s}p(0) + \sigma\int_0^t e^{-\gamma(s-s')}dW\left(s'\right)\right]ds \\
x(t) &= \frac{p(0)}{m}\int_0^t e^{-\gamma s}ds + \frac{\sigma}{m}\int_0^t \left[\int_0^s e^{-\gamma(s-s')}dW\left(s'\right)\right]ds \\
x(t) &= \frac{p(0)}{m}\int_0^t e^{-\gamma s}ds + \frac{\sigma}{m}\int_0^t \left[\int_0^s e^{-\gamma(s-s')}ds'\right]dW(s) \\
x(t) &= \frac{1}{m\gamma}\left(1-e^{-\gamma t}\right)p(0) + \frac{\sigma}{m\gamma}\int_0^t \left(1-e^{-\gamma s}\right)dW(s)
\end{aligned} \tag{5.30}$$

This is the complete solution for x(t). We see from this that the probability for x(t) is a Gaussian. We can now easily calculate the mean and variance, which are,

$$\langle x(t)\rangle = \left\langle \frac{1}{m\gamma}\left(1-e^{-\gamma t}\right)p(0)\right\rangle + \left\langle \frac{\sigma}{m\gamma}\int_0^t \left(1-e^{-\gamma s}\right)dW(s)\right\rangle$$

For a Weiner process, we know that $\langle W\rangle = 0$. Thus,

$$\langle x(t)\rangle = \left\langle \frac{1}{m\gamma}\left(1-e^{-\gamma t}\right)p(0)\right\rangle$$

All of these are constants:

$$\langle x(t)\rangle = \frac{1}{m\gamma}\left(1-e^{-\gamma t}\right)p(0) \tag{5.31}$$

As for variance:

$$\mathbb{V}(x(t)) = \mathbb{V}\left(\frac{1}{m\gamma}\left(1 - e^{-\gamma t}\right) p(0)\right) + \mathbb{V}\left(\frac{\sigma}{m\gamma}\int_0^t \left(1 - e^{-\gamma s}\right) dW(s)\right)$$

Since $\frac{1}{m\gamma}\left(1 - e^{-\gamma t}\right) p(0)$ is deterministic, its variance reduces to zero. For the left most term, we apply Ito's rule,

$$\mathbb{V}(x(t)) = \frac{\sigma^2}{(m\gamma)^2}\int_0^t \left(1 - e^{-\gamma s}\right)^2 ds = \frac{\sigma^2 t}{(m\gamma)^2} + \frac{\sigma^2}{2m^2\gamma^3}\left[4e^{-\gamma t} - e^{-2\gamma t} - 3\right]$$

Langevin recognized that the damping (or decay) rate γ is very high—significantly faster than the time-resolution at which particles could be observed experimentally, especially in the early 20th century. This implies that $\gamma t \gg 1$ during typical observations. Under these conditions, the variance of the particle's position $x(t)$ simplifies to the following approximation:

$$\begin{aligned}
\mathbb{V}(x(t)) &\approx \frac{\sigma^2 t}{(m\gamma)^2} + \frac{\sigma^2}{2m^2\gamma^3} \\
&= \frac{\sigma^2}{(m\gamma)^2}\left(t - \frac{3}{2\gamma}\right) \\
&\approx \frac{\sigma^2 t}{(m\gamma)^2} \\
&= \left(\frac{kT}{3\pi\eta d}\right) t
\end{aligned} \tag{5.32}$$

This result shows that for $t\gamma \gg 1$, the particle's position variance becomes directly proportional to time. This time-dependent variance mirrors the behavior of Wiener noise, which is why Wiener noise is often equated with Brownian motion.

To verify the accuracy of this model for Brownian motion, experiments can be performed by tracking the displacement of a particle (e.g., a pollen grain) over time. By observing the particle's movement during fixed time intervals T, calculating the average squared displacement over multiple trials, and comparing this variance, researchers can confirm the proportionality to time. This approach was successfully used in 1910 by Smoluchowski and others to demonstrate the linear scaling of variance with time. Modern experiments have confirmed this behavior, though deviations from Wiener noise are observed at very short timescales.

The formal definition for $x(t)$ to behave like a Wiener process, the following must hold:

$$dx = \alpha dW, \quad \text{or equivalently} \quad \frac{dx}{dt} \approx \eta(t)$$

for some constant α.

Since $\frac{dx}{dt} = \frac{p}{m}$, the particle's momentum $p(t)$ effectively acts as the noise term $\xi(t)$. This equivalence arises because the autocorrelation function of $p(t)$ decays at a rate

determined by γ. When γ is sufficiently large, the autocorrelation of $p(t)$ becomes sharply peaked, closely resembling the delta-function autocorrelation of $\xi(t)$.

It's important to note that when we describe γ as "large," this is always in comparison to some relevant timescale. In this context, the comparison is made to the observation resolution δt—the time interval over which the process is sampled. If the time resolution δt satisfies

$$\delta t \gg \frac{1}{\gamma},$$

the process $x(t)$ becomes indistinguishable from a Wiener process during observations. Therefore, stating that γ is large means it significantly exceeds $\frac{1}{\delta t}$, ensuring that $p(t)$ appears statistically equivalent to $\xi(t)$.

The power spectrum of Brownian Motion

We know that white noise has a constant power spectrum over all frequencies as indicated by (4.22). The position $x(t)$ of a Brownian particle is the integral of the white noise impacts:

$$x(t) = \int_0^t \eta(t')\, dt'. \tag{5.33}$$

The Fourier Transform of $x(t)$ is:

$$\tilde{x}(f) = \int_{-\infty}^{\infty} x(t) e^{-i2\pi ft}\, dt = \int_{-\infty}^{\infty} \left(\int_0^t \eta(t')\, dt' \right) e^{-i2\pi ft}\, dt. \tag{5.34}$$

In frequency space, integrating $\eta(t)$ corresponds to dividing by $i2\pi f$:

$$\tilde{x}(f) = \frac{\tilde{\eta}(f)}{i2\pi f}.$$

The power spectrum of $x(t)$, $S_x(f)$, is then:

$$S_x(f) = \langle |\tilde{x}(f)|^2 \rangle = \frac{S_\eta(f)}{(2\pi f)^2}. \tag{5.35}$$

Since $S_\eta(f)$ is constant, we have:

$$S_x(f) \propto \frac{1}{f^2}. \tag{5.36}$$

This $\frac{1}{f^2}$ decay is a defining characteristic of Brownian motion in the frequency domain.

6 Fokker-Planck Equation:

Until now, we have been directly focusing on the coordinates x of a single particle undergoing Brownian motion. However, in many cases, we are dealing with an ensemble of particles, making it highly unfeasible to track each particle individually. In such scenarios, it is more fruitful to consider the probability distribution $P(x,t)$ —the likelihood of finding a particle at position x at time t. This distribution is governed by the **Fokker-Planck equation** which will be the subject of interest in the following sections.

6.1 Deriving the Fokker-Planck Equation

Given a stochastic process x(t) with the following Ito differential equation

$$dx = f(x,t)dt + g(x,t)dW \tag{6.1}$$

let us derive the Fokker-Planck equation. To do so, we first calculate the differential equation for the mean value of an arbitrary function h(x) of the stochastic variable x. Firstly, applying Ito's rule for h(x),

$$dh(t,x) = \left(\frac{\partial h}{\partial x}\right) d\boldsymbol{x} + \left(\frac{dh}{\partial t}\right) dt + \frac{1}{2}\left(\frac{d^2h}{dx^2}\right)(d\boldsymbol{x})^2$$

The function $h(x)$ does not depend upon time,

$$dh(t,x) = \left(\frac{\partial h}{\partial x}\right) d\boldsymbol{x} + \frac{1}{2}\left(\frac{d^2h}{d\boldsymbol{x}^2}\right)(d\boldsymbol{x})^2$$

Plugging equation (6.1):

$$dh(t,x) = \left(\frac{\partial h}{\partial x}\right)[f(x,t)dt + g(x,t)dW] + \frac{1}{2}\left(\frac{d^2h}{dx^2}\right)(f(x,t)dt + g(x,t)dW)^2$$

The cross terms would not survive and using Ito's relation $dW^2 = \mathrm{dt}$, obtain:

$$dh(t,\boldsymbol{x}) = \left(\frac{\partial h}{\partial x}\right)[f(x,t)dt + g(x,t)dW] + \frac{1}{2}\left(\frac{d^2h}{dx^2}\right)g^2(x,t)dt$$
$$dh(t,\boldsymbol{x}) = \left(\frac{\partial h}{\partial x}\right)f(x,t)dt + \frac{1}{2}\left(\frac{d^2h}{dx^2}\right)g^2(x,t)dt + \left(\frac{\partial h}{\partial x}\right)g(x,t)dW$$

Taking average on both sides, and recalling that $\langle W\rangle = 0$, we have:

$$d\langle h\rangle = \left\langle \left(\frac{\partial h}{\partial x}\right)f(x,t)dt \right\rangle + \frac{1}{2}\left\langle \left(\frac{d^2h}{dx^2}\right)g^2(x,t)dt \right\rangle$$
$$\frac{d\langle h\rangle}{dt} = \int_{-\infty}^{\infty}\left[f(x,t)\left(\frac{\partial h}{\partial x}\right) + \frac{1}{2}g^2(x,t)\left(\frac{d^2h}{dx^2}\right)\right]P(x,t)dx \tag{6.2}$$

We now perform integration by parts for both of the terms. Once for the first one and twice for the latter.

Integration by parts for the first term

Consider first the term $\int_{-\infty}^{\infty} f(x,t)\left(\frac{\partial h}{\partial x}\right) P(x,t)dx$. We perform integration by parts:

$$u = P(x,t), \quad dv = \frac{dh}{dx}dx.$$

Then,

$$du = \frac{\partial P(x,t)}{\partial x}dx, \quad v = h(x).$$

Such that:

$$\int_{-\infty}^{\infty} f\left(\frac{\partial h}{\partial x}\right) P dx = [h(x)f(x,t)P(x,t)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} h(x)\frac{\partial}{\partial x}(f(x,t)P(x,t))dx$$

Note that a valid PMF $P(x,t)$ must converge to zero at infinity. Thus, the first term in above expression should be zero and it should simplify to the following:

$$\int_{-\infty}^{\infty} f(x,t)\left(\frac{\partial h}{\partial x}\right) P(x,t)dx = -\int_{-\infty}^{\infty} h(x)\frac{\partial}{\partial x}(f(x,t)P(x,t))dx \tag{6.3}$$

Integration by parts for the second term

Consider the second term $\int_{-\infty}^{\infty} g^2(x,t)\left(\frac{d^2h}{dx^2}\right) P(x,t)dx$ now. We perform integration by parts twice on this. The first time:

$$u = P(x,t)g^2(x,t), \quad dv = \frac{d^2h}{dx^2}.$$

Then,

$$du = \frac{\partial}{\partial x}\left[P(x,t)g^2(x,t)\right], \quad v = \frac{dh}{dx}.$$

Applying the integration by parts formula:

$$\int_{-\infty}^{\infty} g^2(x,t)\left(\frac{d^2h}{dx^2}\right) P(x,t)dx = -\int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left[P(x,t)g^2(x,t)\right]\frac{dh}{dx}dx \tag{6.4}$$

Where the boundary term $\left[P(x,t)g^2(x,t)\frac{dh}{dx}\right]_{-\infty}^{\infty}$ once again goes to 0 . Applying integration on (6.4) again.

We have:

$$u = \frac{\partial}{\partial x}\left[P(x,t)g^2(x,t)\right], \quad dv = \frac{dh}{dx}.$$

Then,

$$du = \frac{\partial^2}{\partial x^2}\left[P(x,t)g^2(x,t)\right], \quad v = h(x).$$

Applying the integration by parts formula:

$$\int_{-\infty}^{\infty} g^2(x,t)\left(\frac{d^2h}{dx^2}\right)P(x,t)dx = \int_{-\infty}^{\infty} h(x)\frac{\partial^2}{\partial x^2}\left[P(x,t)g^2(x,t)\right]dx \tag{6.5}$$

Where the boundary term $\left[h(x)\frac{\partial}{\partial x}\left[P(x,t)g^2(x,t)\right]\right]_{-\infty}^{\infty}$ once again goes to 0.

In terms of (6.5) and (6.3), the original equation (6.2) becomes:

$$\frac{d\langle h\rangle}{dt} = \int_{-\infty}^{\infty} -h(x)\frac{\partial}{\partial x}\left(f(x,t)P(x,t)\right)dx + \int_{-\infty}^{\infty}\frac{1}{2}h(x)\frac{\partial^2}{\partial x^2}\left[P(x,t)g^2(x,t)\right]dx$$

$$\frac{d\langle h\rangle}{dt} = \int_{-\infty}^{\infty} h(x)\left\{-\frac{\partial}{\partial x}\left(f(x,t)P(x,t)\right) + \frac{\partial^2}{\partial x^2}\left[P(x,t)g^2(x,t)\right]\right\}dx \tag{6.6}$$

The mean of f is given by,

$$\langle h\rangle = \int_{-\infty}^{\infty} h(x)P(x,t)dx \tag{6.7}$$

Thus, the derivative of mean can be written as,

$$\frac{d\langle h\rangle}{dt} = \frac{d}{dt}\int_{-\infty}^{\infty} h(x)P(x,t)dx = \int_{-\infty}^{\infty} h(x)\frac{\partial}{\partial t}P(x,t)dx \tag{6.8}$$

Equation (6.8) and (6.6) together yield:

$$\int_{-\infty}^{\infty} h(x)\frac{\partial}{\partial t}P(x,t)dx = \int_{-\infty}^{\infty} h(x)\Bigg\{-\frac{\partial}{\partial x}\left[f(x,t)P(x,t)\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[P(x,t)g^2(x,t)\right]\Bigg\}dx. \tag{6.9}$$

Equation (6.9) should hold for any h(x). Thus, the Fokker-Plank Equation becomes the following,

$$\frac{\partial}{\partial t}P(x,t) = -\frac{\partial}{\partial x}(f(x,t)P(x,t)) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[D^2(x,t)P(x,t)\right] \tag{6.10}$$

Where we have defined $D^2(x,t) = g^2(x,t)$. We can write equation (6.10) as a proba-

bility current:

$$\frac{\partial}{\partial t}P(x,t) = -\frac{\partial}{\partial x}\left[f(x,t)P(x,t) + \frac{1}{2}\frac{\partial}{\partial x}\left(D^2(x,t)P(x,t)\right)\right]$$
$$\frac{\partial}{\partial t}P(x,t) = -\frac{\partial}{\partial x}J(x,t). \tag{6.11}$$

Where

$$J(x,t) = f(x,t)P(x,t) + \frac{1}{2}\frac{\partial}{\partial x}\left(D^2(x,t)P(x,t)\right).$$

The relation between P and J , as given by (6.11) implies that J(x) is the probability current and $J(x,t)$ is the rate at which probability is flowing across the point x at time t.

Probability current

To see that $J(x,t)$ corresponds to a probability current, consider the probability that x lies within the narrow interval $[a, a+x]$. This probability is approximately P(a, t)x. Now consider the rate of change of this probability as shown in the figure below:

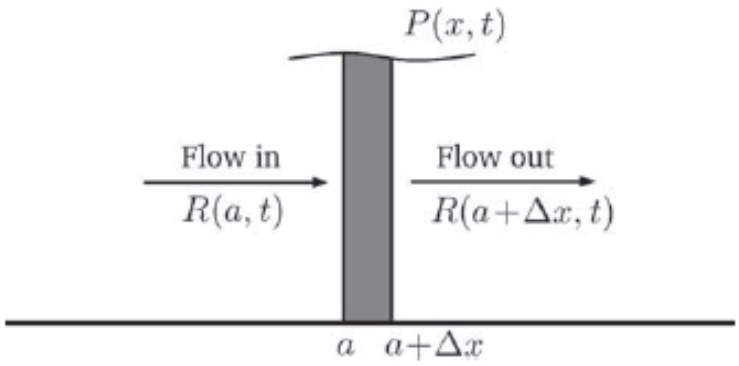


Figure 6: Illustrtion of the Flow across the probability distribuition

This rate of change is given by the difference between the rate at which probability is flowing into the interval from the left, and the rate that it is flowing out from the right. Denoting the rate of flow of probability across the point x at time t as $R(x,t)$, we have:

$$\frac{\partial}{\partial t}[P(a,t)\Delta x] = R(a,t) - R(a+\Delta x,t) \tag{6.12}$$

Dividing both sides by Δx, and taking the limit as $\Delta x \to 0$, we get:

$$\frac{\partial}{\partial t}P(a,t) = -\lim_{\Delta x \to 0}\frac{R(a+\Delta x,t)-R(a,t)}{\Delta x} = -\frac{\partial}{\partial a}R(a,t) \tag{6.13}$$

Comparing this with equation (6.11), we see that J(x, t) is indeed the rate of flow of the probability across the point x .

Boundary Conditions: Absorbing and Reflecting Boundaries:

To solve an FP equation, one may also need to specify the boundary conditions. If x takes values on the entire real line, then this is unnecessary since we know that P tends to zero as $x \to \pm\infty$, and this will be reflected in the initial condition, being the choice for P(x, t) at t = 0. However, if x has some finite domain, say the interval $[a, b]$, then we need to specify what happens at the boundaries a and b . The three most common possibilities are as follows.

1. **Absorbing boundaries**. An absorbing boundary is one in which the particle is removed immediately hits the boundary. This means that the probability that particle is on the boundary is always zero, and this situation is therefore described by the condition

$$\mathrm{P(c,t)} = 0 \tag{6.14}$$

where c is the location of the absorbing boundary.

2. **Reflecting boundaries**. A reflecting boundary is one for which the particle cannot pass through. This means that the probability current must be zero across the boundary, and is therefore given by the condition

$$J(c,t) = 0 \tag{6.15}$$

where c is the location of the reflecting boundary.

3. **Periodic boundaries**. In this case the two ends (boundaries) of the interval are connected together. This means that the particle is moving on a closed loop such as a circle in one dimension, or a torus in two dimensions. In this case, since the two ends describe the same physical location, both the probability density and the probability current must be the same at both ends. This is therefore described by the two conditions

$$\mathrm{P(a,t)} = \mathrm{P(b,t)}, \tag{6.16}$$

$$\mathrm{J(a,t)} = \mathrm{J(b,t)}. \tag{6.17}$$

where the interval in which the particle moves is [a, b].

These three kinds of boundary conditions can also be applied to FP equations in more than one dimension. For reflecting boundaries this means setting to zero the dot product of the vector current with the vector normal to the surface of the boundary.

6.2 Stationary Solution for one dimension:

When the FP equation is one dimensional, one can fairly easily calculate its stationary or steady-state solutions. A stationary solution is defined as one in which P(x, t) does not change with time. The stationary distribution represents the long-term behavior of the stochastic system described by the Fokker-Planck equation. It tells us how the probabilities of the system's states are distributed when the system has reached equilibrium. In physical systems, particularly in thermodynamics and statistical mechanics, the stationary distribution often corresponds to the thermodynamic equilibrium distribution. For example, in systems governed by Boltzmann statistics, the stationary distribution is the Boltzmann distribution, as we will show.

The differential equation that describes the stationary solutions is obtained by setting:

$$\frac{\partial P}{\partial t} = 0 \tag{6.18}$$

Thus, equation (6.11) becomes:

$$\begin{aligned}
&\frac{\partial}{\partial t}P(x,t) = -\frac{\partial}{\partial x}J(x,t) \\
&0 = -\frac{\partial}{\partial x}J(x,t) \\
&0 = -\frac{\partial}{\partial x}[f(x,t)P(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[D^2(x,t)P(x,t)\right] \\
&\frac{d}{dx}\left[D^2(x,t)P(x,t)\right] = 2[f(x,t)P(x,t)]
\end{aligned}$$

Defining a new function $\xi(x) = D(x)P(x)$, we see that this equation is a just a linear differential equation for h :

$$\frac{d\xi}{dx} = \left[\frac{2f(x)}{D(x)}\right]\xi \tag{6.19}$$

The solution is straightforward:

$$P(x) = \frac{1}{\mathcal{N}D(x)}\exp\left[\int_a^x \frac{f(u)}{D(u)}du\right] \tag{6.20}$$

where the particle moves in the interval [a, b], and $\mathcal{N}$ is a constant chosen so that :

$$\mathcal{N} = \int_a^b P(x)dx = 1 \tag{6.21}$$

Notice that this is nothing but the **Maxwellian-Boltzmann distribution** from

statistical mechanics. Thus, we recover our normal physics under equilibrium when $t \to \infty$.

Stationary Solution with Periodic Boundary Conditions

If we have periodic boundary conditions, then J will not necessarily vanish. Neither is J a free parameter, however; as we will see, it is completely determined by the assumption of stationarity. In this case the equation for the stationary solution, P(x), is given by:

$$\frac{d}{dx}[D(x)P(x)] = 2f(x)P(x) - J \tag{6.22}$$

Once again defining $\xi(x) = D(x)P(x)$, this is a linear equation for ξ, but this with a constant driving term:

$$\frac{d\xi}{dx} = \left[\frac{2f(x)}{D(x)}\right]\xi - J \tag{6.23}$$

The solution is:

$$P(x) = \left[\frac{Z(x)}{D(x)}\right]\left\{P(a)\left[\frac{D(a)}{Z(a)}\right] - 2J\int_a^x \frac{du}{Z(u)}\right\}. \tag{6.24}$$

Where we have defined:

$$Z(x) = \exp\left[\int_a^x \frac{f(u)}{D(u)}du\right] \tag{6.25}$$

Now we apply the periodic boundary condition $P(a) = P(b)$ (note that we have already applied the boundary condition on J by making J constant) and this gives:

$$J = \frac{P(a)}{2\int_a^b \frac{du}{Z(u)}}\left[\frac{D(a)}{Z(a)} - \frac{D(b)}{Z(b)}\right] \tag{6.26}$$

The solution is therefore:

$$P(a) = P(a)\frac{\left[\frac{D(b)}{Z(b)}\int_a^x \frac{du}{Z(u)} - \frac{D(a)}{Z(a)}\int_x^b \frac{du}{Z(u)}\right]}{\frac{D(x)}{Z(x)}\int_a^b \frac{du}{Z(u)}} \tag{6.27}$$

6.3 Kolmogorov Backward Equation:

Fokker-Planck Equation as Kolmogorov Forward Equation

Consider an SDE of the form $d\boldsymbol{x} = \mu(X_t, t)\,dt + \sigma(X_t, t)\,d\boldsymbol{w}$, we can define the associated Fokker Plank equation as:

$$\frac{\partial}{\partial t}p(x,t) = -\frac{\partial}{\partial x_t}\left(\mu(x_t,t)\,p(x,t)\right) + \frac{1}{2}\frac{\partial^2}{\partial x_t^2}\left[\sigma^2(x,t)p(x,t)\right]$$

Where $P(x,t)$ is the probability current. We can think of Fokker Plank Equation as a normal distribution being transformed into an arbitrary complex distribution according to the drift and diffusion parameters $\mu(x_t)$ and $\sigma(x_t)$. In fact, if we think in terms of conditional probability densities just like we did way back when considering the conditional distribution to find auto-correlation function, then Fokker Plank Equation basically tells us the evolution:

$$p(x, t \mid x_0, t_0) \tag{6.28}$$

Thus, given (x_0, t_0) FP Equation allows us to find how the system would transition to (x, t).

The Komogorov Backward Equation answers the **opposite question** to FP equation. Namely, how the probability of x_0 at a later point in time changes as we change x_t at an earlier point in time. In completeness, it answers: *'How does the probability of* $\boldsymbol{x_0}$ *at the later point in time t change, as we slowly evolve the probability distribution backwards through time and condition on* $\boldsymbol{x_t}$ *"*

The equation is ideal for problems where the influence of the initial state is crucial such as computing probabilities of reaching a specific future state or understanding how initial conditions impact system behavior. It is defined as following:

$$-\frac{\partial}{\partial t}p(x_0 \mid x_t) = \mu(x_t)\frac{\partial}{\partial x_0}p(x_0 \mid x_t) + \frac{1}{2}\sigma^2(x_t)\frac{\partial^2}{\partial x_0}p(x_0 \mid x_t) \tag{6.29}$$

To motivate the derivation of KBE in a less formal setting, we consider the probability distribution $p(x, t \mid x_0, t_0)$. This distribution must satisfy an important constraint. To see how, consider the example of fish movement and imagine a fish that starts in location x_0 at time t_0 and ends up at location x at time t . We can obtain an expression for $P(x, t \mid x_0, t_0)$ by considering all of the potential locations, x_1, of the fish at some intermediate point in time, t_1. In particular, $P(x, t \mid x_0, t_0)$ can be expressed as the probability of moving from x_0 to x_1 between times t_0 and t_1, and then moving from x_1 to x between times t_1 and t , evaluated over all possible intermediate states, x_1. We can write this logical statement mathematically as,

$$p\left(x,t \mid x_0,t_0\right) = \int p\left(x,t \mid x_1,t_1\right) p\left(x_1,t_1 \mid x_0,t_0\right) dx_1 \tag{6.30}$$

where this integral (and those that follow) is evaluated over the range of possible values of the random variable x_1. Equation (6.30) is known as the **Chapman-Kolmogorov equation**. Using this equation, we make headway into deriving (6.30). Consider an intermediate point in time, t_1, between the present time t and the initial time point t_0 but one that is very close to t_0. That is, $t_1 = t_0 + \Delta t$ where Δt is very small. The crux of the derivation revolves around the assumption that the change in the random variable X(t) over the short time period Δt is small enough that we can use a Taylor series with respect to this change. In particular, after a small amount of time Δt elapses, the value of X is assumed to change by a small amount Δx. Thus we can write the value of X at time t_1 as $x_1 = x_0 + \Delta x$.

Our goal is to derive an expression for the derivative, $\frac{\partial P}{\partial t_0}$. We start by using the definition for the derivative:

$$\frac{\partial p\left(x,t \mid x_0,t_0\right)}{\partial t_0} = \lim_{\Delta t \to 0} \frac{p\left(x,t \mid x_0,t_0+\Delta t\right) - p\left(x,t \mid x_0,t_0\right)}{\Delta t} \tag{6.31}$$

Thus to obtain the desired expression we must obtain an expression for the ratio:

$$\frac{p\left(x,t \mid x_0,t_0+\Delta t\right) - p\left(x,t \mid x_0,t_0\right)}{\Delta t} \tag{6.32}$$

First, we can replace $p\left(x,t \mid x_0,t_0\right)$ in this ratio with the Chapman-Kolmogorov equation (6.30). In addition, we know that $\int p\left(x_1,t_1 \mid x_0,t_0\right) dx_1 = 1$ because the fish must be located somewhere at time t_1. Consequently, we are free to replace $p\left(x,t \mid x_0,t_0+\Delta t\right)$ in (6.21) with:

$$p\left(x,t \mid x_0,t_0+\Delta t\right) = p\left(x,t \mid x_0,t_0+\Delta t\right) \int p\left(x_1,t_1 \mid x_0,t_0\right) dx_1 \tag{6.33}$$

Plugging (6.20) and (6.22) into (6.23):

$$\frac{1}{\Delta t}\left[p\left(x,t \mid x_0,t_0+\Delta t\right) \int p\left(x_1,t_1 \mid x_0,t_0\right) dx_1 - \int p\left(x,t \mid x_1,t_1\right) p\left(x_1,t_1 \mid x_0,t_0\right) dx_1\right] \tag{6.34}$$

Let $t_1 = t_0 + \Delta t$, move $p\left(x,t \mid x_0,t_1\right)$ inside the integral and take $p\left(x_1,t_1 \mid x_0,t_0\right)$ common:

$$\frac{1}{\Delta t}\left[p(x,t \mid x_0,t_1) \int p(x_1,t_1 \mid x_0,t_0) dx_1 - \int p(x,t \mid x_1,t_1) p(x_1,t_1 \mid x_0,t_0) dx_1\right]$$
$$= \frac{1}{\Delta t} \int \Big[p(x,t \mid x_0,t_1) - p(x,t \mid x_1,t_1)\Big] p(x_1,t_1 \mid x_0,t_0) dx_1$$

$$= \frac{1}{\Delta t} \int \left\{ -\Delta x \frac{\partial p(x,t \mid x_0,t_1)}{\partial x_0} - \frac{\Delta x^2}{2} \frac{\partial^2 p(x,t \mid x_0,t_1)}{\partial x_0^2} - O(\Delta x^3) \right\} p(x_1,t_1 \mid x_0,t_0) dx_1. \tag{6.35}$$

Where we used the fact that $x_1 = x_0 + \Delta x$ to take the Taylor series of the term within the square brackets. At this point, we drop the higher-order terms $O\left(\Delta x^3\right)$. Replacing Δx, with $x_1 - x_0$ and factoring out terms that do not depend on x_1 leaves:

$$\frac{1}{\Delta t} \int -\left(x_1 - x_0\right) \frac{\partial p\left(x,t \mid x_0,t_1\right)}{\partial x_0} dx_1 - \frac{1}{2\Delta t} \frac{\partial^2 p\left(x,t \mid x_0,t_1\right)}{\partial x_0^2} \int \left(x_1 - x_0\right)^2 \frac{\partial p\left(x,t \mid x_0,t_1\right)}{\partial x_0} dx_1$$

Finally, we put this in the limit by going back to equation (6.21):

$$\begin{aligned} \frac{\partial p\left(x,t \mid x_0,t_0\right)}{\partial t_0} = & \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int -\left(x_1 - x_0\right) \frac{\partial p\left(x,t \mid x_0,t_1\right)}{\partial x_0} dx_1 \\ & - \frac{1}{2\Delta t} \frac{\partial^2 p\left(x,t \mid x_0,t_1\right)}{\partial x_0^2} \int \left(x_1 - x_0\right)^2 \frac{\partial p\left(x,t \mid x_0,t_1\right)}{\partial x_0} dx_1. \end{aligned} \tag{6.36}$$

We replace t_1 with $t_0 + \Delta t$ and take the limit as $\Delta t \to 0$, allowing us to write the above as:

$$\frac{\partial p\left(x,t \mid x_0,t_0\right)}{\partial t_0} = -\mu\left(x_0\right) \frac{\partial p\left(x,t \mid x_0,t_1\right)}{\partial x_0} - \frac{1}{2} \sigma\left(x_0\right) \frac{\partial^2 p\left(x,t \mid x_0,t_1\right)}{\partial x_0^2} \tag{6.37}$$

In this equation, $\mu(x_0)$ and $\sigma^2(x_0)$ reads:

$$\mu\left(x_0\right) = \lim_{\Delta t \to 0} \frac{\int -\left(x_1 - x_0\right) p\left(x_1, t_0 + \Delta t \mid x_0, t_0\right) dx_1}{\Delta t}, \tag{6.38}$$

$$\sigma^2\left(x_0\right) = \lim_{\Delta t \to 0} \frac{\int \left(x_1 - x_0\right)^2 p\left(x_1, t_0 + \Delta t \mid x_0, t_0\right) dx_1}{\Delta t}. \tag{6.39}$$

7 Reverse-Time Stochastic Equation

7.1 Anderson's Formula

Understanding Anderson's Formula

The Anderson Reverse SDE deals with time-reversing a forward SDE of the form:

$$dx = \mu dt + \sigma dw \tag{7.1}$$

We know that this SDE induces a probability distribution $p(x,t)$, describing the likelihood of finding X_t at position x at time t. The evolution of this probability distribution is governed by the FP equation:

$$\frac{\partial p(x,t)}{\partial t} = -\nabla \cdot \big(\mu(x,t)p(x,t)\big) + \frac{1}{2}\nabla^2\big(\sigma(x,t)^2 p(x,t)\big), \tag{7.2}$$

This PDE provides the macroscopic description of the SDE. Our goal is to find an SDE describing the dynamics of X_t backward in time, from $t = T$ to $t = 0$, such that the probability distribution at each time matches the **forward distribution** $p(x,t)$. To reverse an SDE, we consider the process starts at time T with distribution $p(x,T)$ and aim to find the time-reversed dynamics which ensure that the backward process reproduces the same probability distributions as the forward process. Let's denote the reversed process by Y_t, evolving backward in time. Its dynamics are expressed as:

$$dY_t = \tilde{\mu}(Y_t, t)\, dt + \sigma(Y_t, t)\, dW_t^{\text{rev}}, \tag{7.3}$$

where $\tilde{\mu}(Y_t,t)$ is the Drift term of the reversed SDE, $\sigma(Y_t,t)$ is the Diffusion term identical to the forward process and dW_t^{rev} is a Wiener process adapted to the reversed time. The key is to derive $\tilde{f}(x,t)$, the drift of the reverse-time SDE. $\nabla \log p(x,t)$. As we will show, the reverse-time SDE would assume the following form:

$$dx = \Big[\underbrace{\big[\mu(x,t)\big]}_{\text{drift}} - \sigma^2(t)\underbrace{\big[\nabla_x \log p_t(x)\big]}_{\text{score function}}\Big]\, dt + \underbrace{\sigma(t)\, dW_t^{rev}}_{\text{reverse-time diffusion}} \tag{7.4}$$

This result was produced in the paper, **Reverse-time Diffusion Equation Models** in 1982 by BDO Anderson.

In order to prove Anderson's Formula, we are going to be using the Kolmogorov

Backward Equation which we derived before:

$$-\partial_t p\left(x_s \mid x_t\right) = \mu\left(x_t\right) \partial_{x_t} p\left(x_s \mid x_t\right) + \frac{1}{2}\sigma^2\left(x_t\right) \partial_{x_t}^2 p\left(x_s \mid x_t\right) \tag{7.5}$$

Let x_s represents our initial distribution and x_t represents our final distribution. We use Bayes Theorem to make headways:

$$p\left(x_s, x_t\right) = p\left(x_s \mid x_t\right) p\left(x_t\right) \tag{7.6}$$

Consider the Left-hand side. First multiplying both sides of Bayes theorem with minus one and taking the derivative with respect to time t via product rule, we obtain:

$$\begin{aligned} -\partial_t p\left(x_s, x_t\right) &= -\partial_t \left[p\left(x_s \mid x_t\right) p\left(x_t\right)\right] \\ &= -\partial_t p\left(x_s \mid x_t\right) p\left(x_t\right) - p\left(x_s \mid x_t\right) \partial_t p\left(x_t\right). \end{aligned} \tag{7.7}$$

Where $-\partial_t p\left(x_s \mid x_t\right)$ is the Kolmogorov Backward Equation and $\partial_t p\left(x_t\right)$ is the Kolmogorov Forward Equation. Plugging (7.2) and (7.5):

$$\begin{aligned} -\partial_t p(x_s, x_t) = &\left(\mu(x_t)\partial_x p(x_s \mid x_t) + \frac{1}{2}\sigma^2(x_t)\partial_x^2 p(x_s \mid x_t)\right) p(x_t) \\ &+ p(x_s \mid x_t)\left(\partial_x\left[\mu(x_t)p(x_t)\right] + \frac{1}{2}\partial_x^2\left[\sigma^2(x_t)p(x_t)\right]\right). \end{aligned} \tag{7.8}$$

First consider the derivative $\partial_t p(x_s|x_t)$ occurring in the backward Kolmogorov equation:

$$\begin{aligned} \partial_{x_t} p\left(x_s \mid x_t\right) &= \partial_{x_t}\left[\frac{p\left(x_s, x_t\right)}{p\left(x_t\right)}\right], \\ \partial_{x_t} p\left(x_s \mid x_t\right) &= \frac{\partial_{x_t} p\left(x_s, x_t\right) p\left(x_t\right) - p\left(x_s, x_t\right) \partial_{x_t} p\left(x_t\right)}{p^2\left(x_t\right)}, \\ \partial_{x_t} p\left(x_s \mid x_t\right) &= \frac{\partial_{x_t} p\left(x_s, x_t\right)}{p\left(x_t\right)} - \frac{p\left(x_s, x_t\right) \partial_{x_t} p\left(x_t\right)}{p^2\left(x_t\right)}. \end{aligned} \tag{7.9}$$

The next step is to evaluate the derivative of the products in the forward Kolmogorov equation.

$$\partial_{x_t}\left[\mu\left(x_t\right) p\left(x_t\right)\right] = \partial_{x_t}\mu\left(x_t\right) p\left(x_t\right) + \mu\left(x_t\right) \partial_{x_t} p\left(x_t\right), \tag{7.10}$$

$$\partial_{x_t}^2\left[\sigma^2\left(x_t\right) p\left(x_t\right)\right] = \partial_{x_t}^2\sigma^2\left(x_t\right) p\left(x_t\right) + 2\partial_{x_t} p\left(x_t\right) \partial_{x_t}\sigma^2\left(x_t\right) + \sigma^2\left(x_t\right) \partial_{x_t}^2 p\left(x_t\right). \tag{7.11}$$

Plugging equation (7.10) in (7.8):

$$-\partial_t p\left(x_s, x_t\right) = \mu\left(x_t\right) \partial_{x_t} p\left(x_s \mid x_t\right) p\left(x_t\right) + \frac{1}{2}\sigma^2\left(x_t\right) \partial_{x_t}^2 p\left(x_s \mid x_t\right) p\left(x_t\right) + p\left(x_s \mid x_t\right) \partial_{x_t} \mu\left(x_t\right) p\left(x_t\right) + p\left(x_s \mid x_t\right) \mu\left(x_t\right) \partial_{x_t} p\left(x_t\right).$$

Plugging equation (7.9):

$$\begin{aligned} -\partial_t p\left(x_s, x_t\right) = \mu\left(x_t\right) & \left[\frac{\partial_{x_t} p\left(x_s, x_t\right)}{p\left(x_t\right)} - \frac{p\left(x_s, x_t\right) \partial_{x_t} p\left(x_t\right)}{p^2\left(x_t\right)}\right] p\left(x_t\right) \\ & + p\left(x_s \mid x_t\right) \partial_{x_t} \mu\left(x_t\right) p\left(x_t\right) + p\left(x_s \mid x_t\right) \mu\left(x_t\right) \partial_{x_t} p\left(x_t\right) \\ & + \frac{1}{2}\sigma^2\left(x_t\right) \partial_{x_t}^2 p\left(x_s \mid x_t\right) p\left(x_t\right) - \frac{1}{2} p\left(x_s \mid x_t\right) \partial_{x_t}^2 \left[\sigma^2\left(x_t\right) p\left(x_t\right)\right]. \end{aligned}$$

Using $p\left(x_s, x_t\right) = p\left(x_s \mid x_t\right) p\left(x_t\right)$:

$$\begin{aligned} -\partial_t p\left(x_s, x_t\right) = \mu(x_t) & \left[\partial_{x_t} p\left(x_s, x_t\right) - \underline{p\left(x_s \mid x_t\right) \partial_{x_t} p\left(x_t\right)}\right] \\ & + p\left(x_s \mid x_t\right) \partial_{x_t} \mu(x_t) p(x_t) + \underline{p\left(x_s \mid x_t\right) \{\mu(x_t) \partial_{x_t} p(x_t)} \\ & + \frac{1}{2}\sigma^2(x_t) \partial_{x_t}^2 p\left(x_s \mid x_t\right) p(x_t) - \frac{1}{2} p\left(x_s \mid x_t\right) \partial_{x_t}^2 \left[\sigma^2(x_t) p(x_t)\right]. \end{aligned}$$

Where the highlighted terms cancel one another out. Thus, we are left with:

$$\begin{aligned} -\partial_t p\left(x_s, x_t\right) = & \mu(x_t) \partial_{x_t} p(x_s, x_t) + p(x_s \mid x_t) \partial_{x_t} \mu(x_t) p(x_t) \\ & + \frac{1}{2}\sigma^2(x_t) \partial_{x_t}^2 p(x_s \mid x_t) p(x_t) - \frac{1}{2} p(x_s \mid x_t) \partial_{x_t}^2 \left[\sigma^2(x_t) p(x_t)\right]. \end{aligned}$$

Simplifying further, we notice the first two terms arise from a product rule of the term $\partial_{x_t}\left[\mu(x_t) p(x_s, x_t)\right]$:

$$\begin{aligned} -\partial_t p(x_s, x_t) = & \partial_{x_t}\left[\mu(x_t) p(x_s, x_t)\right] \\ & + \frac{1}{2}\sigma^2(x_t) \partial_{x_t}^2 p(x_s \mid x_t) p(x_t) - \frac{1}{2} p(x_s \mid x_t) \partial_{x_t}^2 \left[\sigma^2(x_t) p(x_t)\right]. \end{aligned} \tag{7.12}$$

Now, to make further headways, we observe that the colored terms can arise from the triple product of the term $\partial_{x_t}^2\left[p(x_s|x_t) p(x_t) \sigma^2(x_t)\right]$:

$$\begin{aligned} \frac{1}{2}\partial_{x_t}^2 \left[p(x_s, x_t)\sigma^2(x_t)\right] &= \frac{1}{2}\partial_{x_t}^2 \left[p(x_s \mid x_t) p(x_t) \sigma^2(x_t)\right] \\ &= \frac{1}{2}\partial_{x_t}^2 p(x_s \mid x_t) p(x_t) \sigma^2(x_t) + \partial_{x_t}\left[p(x_t)\sigma^2(x_t)\right] \partial_{x_t} p(x_s \mid x_t) \end{aligned}$$

$$+\frac{1}{2}p(x_s \mid x_t)\partial_{x_t}^2\left[p(x_t)\sigma^2(x_t)\right]. \tag{7.13}$$

We can see from the expansion of the derivative above that we can combine the terms in our derivation if we expand the "center term". Furthermore, we can employ the identity $-\frac{1}{2}X = -X + \frac{1}{2}X$ to obtain:

$$\begin{aligned}-\partial_t p(x_s, x_t) = \partial_{x_t}\left[\mu(x_t)p(x_s, x_t)\right] + \frac{1}{2}\sigma^2(x_t)\partial_{x_t}^2 p(x_s \mid x_t)p(x_t) \\ - \frac{1}{2}p(x_s \mid x_t)\partial_{x_t}^2\left[\sigma^2(x_t)p(x_t)\right] \pm \partial_{x_t}p(x_s \mid x_t)\partial_{x_t}\left[p(x_t)\sigma^2(x_t)\right].\end{aligned}$$

Consider the term $\frac{1}{2}p\left(x_s \mid x_t\right)\partial_{x_t}^2\left[\sigma^2\left(x_t\right)p\left(x_t\right)\right]$. Applying the identity $-\frac{1}{2}X = -X + \frac{1}{2}X$,

$$\begin{aligned}-\partial_t p\left(x_s, x_t\right) = \partial_{x_t}\left[\mu\left(x_t\right)p\left(x_s, x_t\right)\right] + \frac{1}{2}\sigma^2\left(x_t\right)\partial_{x_t}^2 p\left(x_s \mid x_t\right)p\left(x_t\right) \\ -p\left(x_s \mid x_t\right)\partial_{x_t}^2\left[\sigma^2\left(x_t\right)p\left(x_t\right)\right] + \frac{1}{2}p\left(x_s \mid x_t\right)\partial_{x_t}^2\left[\sigma^2\left(x_t\right)p\left(x_t\right)\right] \\ \pm\partial_{x_t}p\left(x_s \mid x_t\right)\partial_{x_t}\left[p\left(x_t\right)\sigma^2\left(x_t\right)\right]\end{aligned}$$

In order for a product rule to arise, we must have the negative sign instead of $\pm$ sign in the term. We therefore obtain:

$$\begin{aligned}-\partial_t p\left(x_s, x_t\right) = \partial_{x_t}\left[\mu\left(x_t\right)p\left(x_s, x_t\right)\right] + \frac{1}{2}\sigma^2\left(x_t\right)\partial_{x_t}^2 p\left(x_s \mid x_t\right)p\left(x_t\right) \\ +\frac{1}{2}p\left(x_s \mid x_t\right)\partial_{x_t}^2\left[\sigma^2\left(x_t\right)p\left(x_t\right)\right] - p\left(x_s \mid x_t\right)\partial_{x_t}^2\left[\sigma^2\left(x_t\right)p\left(x_t\right)\right] \\ -\partial_{x_t}p\left(x_s \mid x_t\right)\partial_{x_t}\left[p\left(x_t\right)\sigma^2\left(x_t\right)\right].\end{aligned}$$

The cyan terms can be seen as arising from the product rule. Similarly, the magenta and purple terms can also be seen as arising from product rule. We thus have,

$$\begin{aligned}-\partial_t p\left(x_s, x_t\right) =& \partial_{x_t}\left[\mu\left(x_t\right)p\left(x_s, x_t\right)\right] + \frac{1}{2}\partial_{x_t}^2\left[p\left(x_s \mid x_t\right)p(x_t)\sigma^2\left(x_t\right)\right] \\ &- \partial_{x_t}\left[p\left(x_s \mid x_t\right)\sigma^2\left(x_t\right)\right]\end{aligned}$$

We now combine the terms with first-order derivatives:

$$\begin{aligned}-\partial_t p\left(x_s, x_t\right) =& \partial_{x_t}\left[\mu\left(x_t\right)p\left(x_s, x_t\right) - p\left(x_s \mid x_t\right)\partial_{x_t}\left[\sigma^2\left(x_t\right)p\left(x_t\right)\right]\right] \\ &+ \frac{1}{2}\partial_{x_t}^2\left[p\left(x_s \mid x_t\right)\sigma^2\left(x_t\right)\right] \\ -\partial_t p\left(x_s, x_t\right) =& \partial_{x_t}\left[p\left(x_s, x_t\right)\left(\mu\left(x_t\right) - \frac{1}{p\left(x_t\right)}\partial_{x_t}\left[\sigma^2\left(x_t\right)p\left(x_t\right)\right]\right)\right] + \frac{1}{2}\partial_{x_t}^2\left[p\left(x_s \mid x_t\right)\sigma^2\left(x_t\right)\right] \\ -\partial_t p\left(x_s, x_t\right) =& -\partial_{x_t}\left[p\left(x_s, x_t\right)\left(-\mu\left(x_t\right) + \frac{1}{p\left(x_t\right)}\partial_{x_t}\left[\sigma^2\left(x_t\right)p\left(x_t\right)\right]\right)\right] +\end{aligned}$$

$$\frac{1}{2}\partial_{x_t}^2\left[p\left(x_s \mid x_t\right)\sigma^2\left(x_t\right)\right] \tag{7.14}$$

In accordance with Leibniz' rule we can marginalize over x_s without interfering with the partial derivative ∂_t to obtain:

$$\begin{aligned}-\partial_t p\left(x_t\right) = -\partial_{x_t}\left[p\left(x_t\right)\left(-\mu\left(x_t\right)+\frac{1}{p\left(x_t\right)}\partial_{x_t}\left[\sigma^2\left(x_t\right)p\left(x_t\right)\right]\right)\right] \\ +\frac{1}{2}\partial_{x_t}^2\left[p\left(x_t\right)\sigma^2\left(x_t\right)\right].\end{aligned}$$

Introducing the time reversal $\tau = 1 - t$ with respect to the integration with respect to the flow of time yields:

$$\begin{aligned}-\partial_t p\left(x_t\right) = -\partial_{x_t}\left[p\left(x_{1-\tau}\right)\left(-\mu\left(x_{1-\tau}\right)+\frac{1}{p\left(x_{1-\tau}\right)}\partial_{x_{1-\tau}}\left[\sigma^2\left(x_{1-\tau}\right)p\left(x_{1-\tau}\right)\right]\right)\right] + \\ \frac{1}{2}\partial_{x_{1-\tau}}^2\left[p\left(x_{1-\tau}\right)\sigma^2\left(x_{1-\tau}\right)\right]\end{aligned}$$

Thus, we obtain:

$$dX_\tau = \left(-\mu\left(x_{1-\tau}\right)+\frac{1}{p\left(x_{1-\tau}\right)}\partial_{x-\tau}\left[\sigma^2\left(x_{1-\tau}\right)p\left(x_{1-\tau}\right)\right]\right)d\tau + \sigma\left(x_{1-\tau}\right)dW_\tau \tag{7.15}$$

Where $\widetilde{W_t}$ is the Wiener process that flows backward in time. Let $\sigma^2\left(x_t\right)$ be constant and independent of x_t. Furthermore, making use of the log-derivative trick, we obtain:

$$\begin{aligned}dX_\tau &= \left(-\mu\left(x_{1-\tau}\right)+\frac{\sigma^2}{p\left(x_{1-\tau}\right)}\partial_{x-\tau}\left[p\left(x_{1-\tau}\right)\right]\right)d\tau + \sigma\left(x_{1-\tau}\right)dW_\tau, \\ dX_\tau &= \left(-\mu\left(x_{1-\tau}\right)+\sigma^2\frac{\partial_{x-\tau}\left[p\left(x_{1-\tau}\right)\right]}{p\left(x_{1-\tau}\right)}\right)d\tau + \sigma\left(x_{1-\tau}\right)dW_\tau, \\ dX_\tau &= \left(-\mu\left(x_{1-\tau}\right)+\sigma^2\partial_{x-\tau}\log p\left(x_{1-\tau}\right)\right)d\tau + \sigma\left(x_{1-\tau}\right)d\widetilde{W_\tau}.\end{aligned} \tag{7.16}$$

Where $\tau = 1 - t$

8 Numerical Methods

8.1 Euler-Maruyama method

The stochastic differential equations (SDEs) that allow for analytical solutions represent only a tiny subset of all possible equations. For most SDEs computational approaches such as numerical simulations, are required to solve them. When solving

an equation of the form:

$$d\boldsymbol{x} = f(x,t)dt + g(x,t)d\boldsymbol{W} \tag{8.1}$$

a numerical approach begins by assigning an initial value to x. The differential equation is then approximated using the discretized form:

$$\Delta\boldsymbol{x} = f(x,t)\Delta t + g(x,t)\Delta\boldsymbol{W} \tag{8.2}$$

Where Δt is a small, fixed time increment, and ΔW is a Gaussian random variable with a mean of zero and variance Δt. For each time step Δt, Δx is computed by generating a random value for ΔW, then added to x. This process is repeated, iteratively updating x at each time step.

This simulation yields an approximation of a single sample path for x. To estimate the probability density of x at a specific time T, the simulation is repeated multiple times, with each iteration using a different set of random values for ΔW. By aggregating the resulting sample paths at T, a histogram of the outcomes can be constructed, providing an approximate probability distribution of $x(T)$. Additionally, the mean and variance of $x(T)$ can be approximated by computing the mean and variance of the generated samples.

This method extends to vector SDEs involving a set of variables $\boldsymbol{x} = (x_1, \ldots, x_N)$ and a vector of independent noise increments $\boldsymbol{dW} = (dW_1, \ldots, dW_M)$:

$$d\boldsymbol{x} = f(\boldsymbol{x},t)\,dt + G(\boldsymbol{x},t)\,d\boldsymbol{W}, \tag{8.3}$$

where G is an $N \times M$ matrix. As before, Δt replaces dt, and $\Delta\boldsymbol{W} = (\Delta W_1, \ldots, \Delta W_M)$ substitutes for $d\boldsymbol{W}$ in the approximation of $\Delta\boldsymbol{x}$. This numerical approach is analogous to Euler's method for deterministic equations.

Simulating Brownian Motion

Previously, we came across Geometric Brownian Motion. The SDE for which is the following:

$$dX_t = \mu X_t\,dt + \sigma X_t\,dW_t,$$

where μ is the drift and σ is the volatility. For this SDE, we found the analytical solution to be:

$$X_t = X_0 \exp\left[\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W_t\right]$$

We now set the value of initial parameters $\mu = 0.1$, $\sigma = 0.1$ and $X_0 = 1$ and approximate the solution using Euler-Murayama as:

$$X_{i+1} = X_i + \mu X_i \Delta t + \sigma X_i \Delta W_i, \tag{8.4}$$

where $\Delta W_i \sim \mathcal{N}(0, \sqrt{\Delta t})$ is the Brownian increment. The procedure we follow is:

1. Initialize $X_0 = 1.0$.
2. For $i = 1, 2, \ldots, N$:
   (a) Compute $\Delta W_i \sim \mathcal{N}(0, \sqrt{\Delta t})$.
   (b) Update $X_{i+1} = X_i + \mu X_i \Delta t + \sigma X_i \Delta W_i$.

This yields the following simulation:

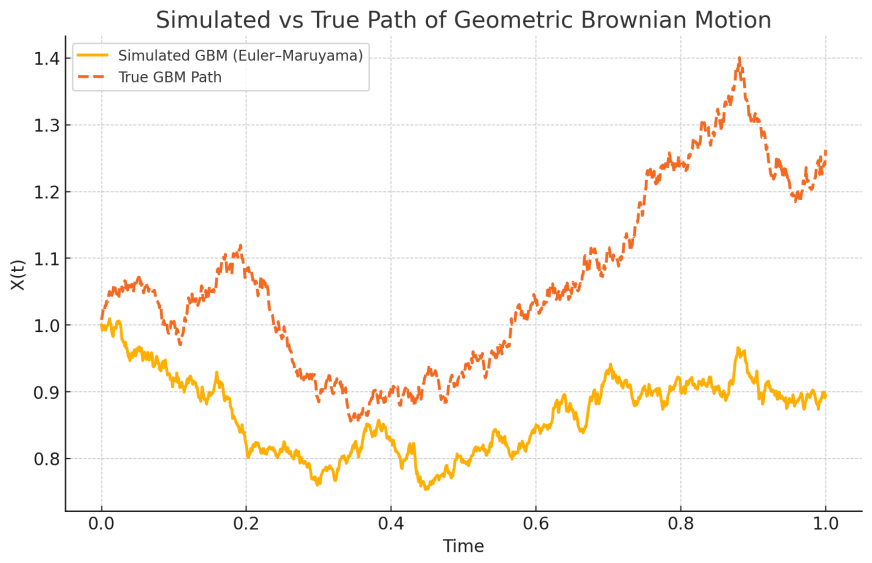


Figure 7: Simulation of GBM using Euler-Murayama Method

Where we had set the interval to $\Delta t = 0.001$ to obtain the solution. The python code can be found at Section A.1.1

The accuracy of the generated sample paths and quantities of interest such as the mean and variance depends on the time step Δt as well as the values of $\boldsymbol{x}$, f, and g throughout the simulation. Smaller time steps result in higher accuracy, with the simulated paths converging to true paths as Δt approaches zero.

One method to assess the accuracy of a simulation with a specific time-step size is to repeat the simulation using half the original time-step. If the results of both simulations are very similar, the accuracy of the initial simulation can be considered approximately equal to the difference between the two results. This is based on the expectation that halving the time-step again would cause an even smaller change in the outcome than the first reduction.

The process of halving the time-step, Δt, in a simulation warrants closer examina-

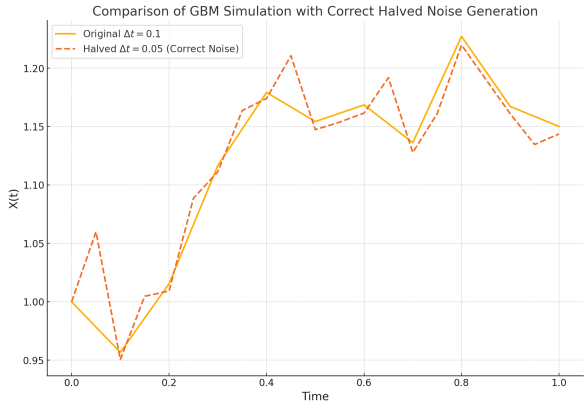
tion. A given sample path is determined by a specific realization of noise, represented by a sequence of random variables ΔW, which are generated for the simulation. If the simulation consists of N time-steps, each noise increment can be labeled as ΔW_n for $n = 0, 1, \ldots, N-1$. To halve the time-step and approximate the same sample path, it is necessary to generate $2N$ Gaussian random variables, denoted by $\Delta\widetilde{W}_m$, that are consistent with the original set of N random variables ΔW_n.

Specifically, the sum of the first two increments in the new simulation, $\Delta\widetilde{W}_0 + \Delta\widetilde{W}_1$, must match the first noise increment ΔW_0 from the original simulation. This ensures that the total stochastic increment over the same time interval is identical for both simulations, thereby maintaining consistency in the noise realization. Similarly, this condition must hold for subsequent pairs of increments, such as $\Delta\widetilde{W}_2 + \Delta\widetilde{W}_3$, and so on.

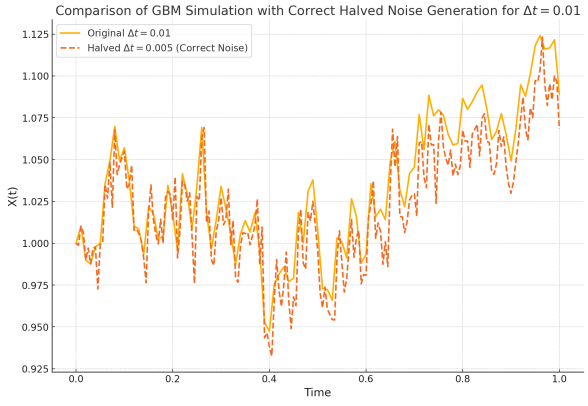
The requirement can be expressed as:

$$\Delta\widetilde{W}_{2n} + \Delta\widetilde{W}_{2n+1} = \Delta W_n, \quad n = 0, 1, \ldots, N-1 \tag{8.5}$$

The figure below shows this process for GBM for $\Delta t = 0.1$ and $\Delta t = 0.01$. The discrepancy between the two paths is more noticeable when Δt is larger.



(a) Simulation for $\Delta t = 0.1$



(b) Simulation for $\Delta t = 0.01$

Fortunately it is very easy to generate a set of $\Delta\widetilde{W}_{2n}$ for which 8.5 is true. All one has to do is generate N random numbers r_n with mean zero and variance $\frac{\Delta t}{2}$, and then

set:

$$\begin{cases} \Delta\widetilde{W}_{2n} = r_n \\ \Delta\widetilde{W}_{2n+1} = \Delta W_n - r_n \end{cases} \tag{8.6}$$

The above procedure allows one to perform two simulations of the same sample path for an SDE with different time-steps. If the difference between the final values of x for the two simulations are too large, then one can halve the time-step again and perform another simulation. One stops when the process of halving the time-step changes the final value of x by an amount that is considered to be small enough for the given application.

By repeatedly halving the time-step, one can also determine how rapidly the simulation converges to the true value of $x(\mathrm{T})$. The faster the convergence the better, and different numerical methods have different rates of convergence. The simple Euler method that we described above has the slowest rate of convergence.

Taylor Expansion of Stochastic Series

For the general SDE of the form:

$$dX_t = f(X_t)\,dt + g(X_t)\,dW_t,$$

The solution $X_{t_{n+1}}$ can be expanded as a stochastic Taylor series:

$$X_{t_{n+1}} = X_{t_n} + f(X_{t_n})\Delta t + g(X_{t_n})\Delta W_n + \frac{\partial g}{\partial x} g(X_{t_n}) \int_{t_n}^{t_{n+1}} (W_s - W_{t_n})\,dW_s + \cdots \tag{8.7}$$

The Euler–Maruyama method truncates the Taylor series after the first two terms:

$$X_{t_{n+1}} \approx X_{t_n} + f(X_{t_n})\Delta t + g(X_{t_n})\Delta W_n. \tag{8.8}$$

The Order of the Method: The accuracy of a numerical method is referred to as the order of the method. Adopting this lingo, Euler-Muruyama method is what we call a **half-order** method. This is because the global error in the Euler–Maruyama method decreases proportionally to $\sqrt{(\Delta t)}$. This is in contrast to deterministic methods (e.g., Euler's method for ODEs), where the error decreases proportionally to Δt.

To understand this, consider the Taylor Expansion of the exact solution. We see that the next higher-order term that Euler-Maruyama omits is the integral:

$$\int_{t_n}^{t_{n+1}} (W_s - W_{t_n})\,dW_s. \tag{8.9}$$

This is stochastic integral over the interval $[t_n, t_{n+1}]$ involving the deviation of W_s from W_{t_n}. Since W_s scales as $\sqrt{s}$, the deviation $(W_s - W_{t_n})$ grows with the square root of the time-step Δt. The integral adds a second scaling factor because it involves

another stochastic increment dW_s which also scales as $\sqrt{\Delta t}$.

Combining these effects:

$$\int_{t_n}^{t_{n+1}} (W_s - W_{t_n})\, dW_s \propto (\Delta t)^{3/2}. \tag{8.10}$$

This represents the **Local Error**. To obtain the **Global Error**, which is the cumulative error over the entire simulation, we consider the time-steps $N = T/\Delta t$:

$$\text{Global error} \propto N \cdot (\Delta t)^{3/2} = \frac{T}{\Delta t} \cdot (\Delta t)^{3/2} \propto (\Delta t)^{1/2} \tag{8.11}$$

Thus the global error scales as $(\Delta t)^{1/2}$ for the Euler-Maruyama Method.

8.2 Beyond Euler-Maruyama

Beyond Euler method, **Milstein** method considers the first three-terms in the Taylor Expansion of Stochastic Integrals, thereby approximating the SDE using:

$$\Delta x = f(x,t)\Delta t + g(x,t)\Delta W + \frac{g(x,t)}{2}\frac{\partial g(x,t)}{\partial x}\left[(\Delta W)^2 - \Delta t\right] \tag{8.12}$$

The third term is called the **The Lévy area term**. The global error of this method is 1 with the error scaling by Δt. While the Milstein method achieves first-order strong convergence, a key disadvantage is the need to compute the derivative $g'(x, y)$. This can be computationally expensive or impractical for certain functions

To address this limitation, the **Milstein-Platen** method replaces the derivative with an approximation. This approximation avoids explicit differentiation and instead uses finite differences. The first-order approximation of the derivative term is:

$$g(x,t)\frac{\partial}{\partial x}g(x,t) \approx \frac{1}{\sqrt{\Delta t}}[g(q,t) - g(x,t)] \tag{8.13}$$

Where q is a predictor value defined as:

$$q = x + f(x,t)\Delta t + g(x,t)\sqrt{\Delta \mathrm{t}} \tag{8.14}$$

Substituting this into Milstien's method for a single variable, we obtain the Milstien-Platen method for a single variable:

$$\Delta x = f\Delta t + g\Delta W + \frac{1}{2\sqrt{\Delta t}}\left[g(q,t) - g(x,t)\left[(\Delta W)^2 - \Delta t\right]\right] \tag{8.15}$$

9 Diffusion Models:

In thermodynamics, diffusion describes the spontaneous flow of particles from regions of higher density to regions of lower density. This process is driven by entropy maximization. The concept can be analogously extended to the statistical domain where diffusion refers to the transformation of a complex probability distribution, p_{complex} on $\mathbb{R}^d$, into a simpler distribution, p_{prior}, over the same space. Formally, this process can be represented as a transformation:

$$\Phi : \mathbb{R}^d \to \mathbb{R}^d \tag{9.1}$$

such that:

$$x_0 \sim p_{\text{complex}} \quad \Longrightarrow \quad \Phi(x_0) \sim p_{\text{prior}}. \tag{9.2}$$

Here, the transformation Φ acts analogously to the physical diffusion process, redistributing the "density" of samples in $\mathbb{R}^d$ to achieve a target equilibrium characterized by p_{prior}. Here, ϕ is a stochastic process whose evolution can be described by Fokker–Planck dynamics in which the system evolves towards a steady-state distribution (analogous to p_{prior}).

We can think of this concept in terms of the *Kolmogorov Forward Equation* which is rooted in Markov chains and their stationary distributions. It states that repeated applications of a transition kernel $q\left(\vec{x} \mid \vec{x}'\right)$ to samples from any initial distribution will eventually produce samples from the target distribution $p_{\text{prior}}(\vec{x})$ provided the following condition holds:

$$p_{\text{prior}}(x) = \int q\left(x \mid x'\right) p_{\text{complex}}\left(x'\right) dx'. \tag{9.3}$$

Here, $q\left(x \mid x'\right)$ defines the Markov transition kernel, and $p_{\text{prior}}(x)$ emerges as the stationary distribution of the chain. This ensures that repeated applications of the kernel guide the samples from the complex initial distribution p_{complex} towards the desired simple distribution p_{prior}.

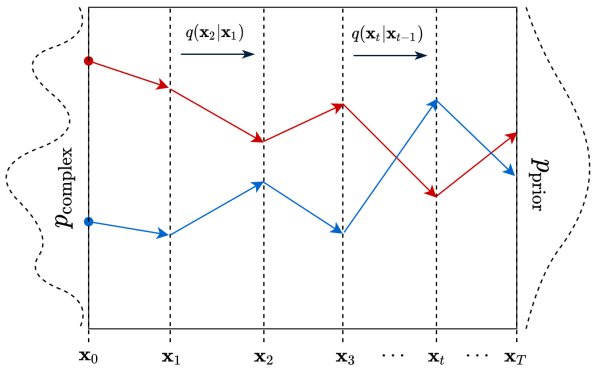


Figure 9: Diffusion as repeated application of Kernel, Figure taken from Ayan Das's Blogspot

We use any data distribution (let's denote it as p_{data}) of our choice as the complex initial density. This leads to the forward diffusion process:

$$x_0 \sim p_{\text{data}} \quad \Rightarrow \quad x_T = \Phi(x_0) \sim \mathcal{N}(0, \mathbb{I}) \tag{9.4}$$

where this process "destructures" the data, turning it into an isotropic Gaussian. However, this forward process by itself is not particularly useful. What is useful is the reverse process: starting from isotropic Gaussian noise and transforming it back into p_{data}—this constitutes generative modeling. Since the forward process is fixed (non-parametric) and guaranteed to exist, it is possible to invert it. Once inverted, it can be used as a generative model as follows:

$$x_T = \Phi(x_0) \sim \mathcal{N}(0, \mathbb{I}) \quad \Rightarrow \quad \Phi^{-1}(x_T) \sim p_{\text{data}}. \tag{9.5}$$

We can think of the diffusion process in terms of *manifold learning*. The manifold hypothesis states that real-world data (e.g., images, audio, text) often lies on or near a lower-dimensional manifold embedded within a higher-dimensional ambient space. Thus, p_{data} is not uniformly distributed across the full space and instead is concentrated within a structure subspace.

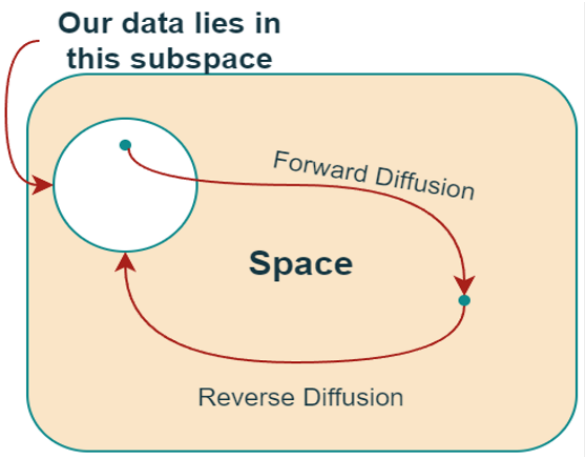


Figure 10: Diffusion as capturing the manifold of the data, Figure taken from a Blogspot

In the figure, the subspace (white circle) represents the lower-dimensional manifold where the data distribution p_{data} resides while the ambient space (beige background) represents the full high-dimensional space. At the starting point x_0, the data is highly structured and confined to the manifold. The forward diffusion process destroys the manifold's structure, progressively injecting noise into the data representation. This eventually resulting in a representation x_T that is isotropic and unstructured. The reverse diffusion process inverts this destructuring process and recovers the structured data distribution confined to the lower-dimensional manifold.

9.1 Forward Diffusion Process:

The following formulation is based upon the paper, "Denoising Diffusion Probabilistic Models" (or DDPM for short). We begin by indexing the steps in the process. The convention is to define the process to have T time steps, where T is some big integer (for example, in the DDPM they use T = 1000). The first step, where we have the original image without any noise, is at t = 0, and we mark that image as x_0 where x_0 is a random vector. Each component of x_0 (e.g. each pixel value) is a random variable. In the following steps, we add noise to the image, so for a large enough T, we will get a complete noise.

The original images (those without noise) come from a distribution of "real data", so we can write it mathematically as:

$$x_0 \sim q(x) \tag{9.6}$$

Where $q(x)$ involves natural images from a real-world dataset.
In the forward diffusion process at a time t, we want to sample a new noisy image based on the noisy image at the previous step $t-1$. That is,

$$x_t \sim q\left(x_t \mid x_{t-1}\right) \tag{9.7}$$

Let us define a forward diffusion process in which we add small amount of Gaussian noise to the samples in T steps, producing a sequence of noisy samples $\vec{x}_1, \ldots, \vec{x}_T$. The step sizes are controlled by a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$

$$\vec{x}_t \sim q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbb{I}\right) \tag{9.8}$$

A few remarks on Notation and Schedular β_t

If you're unfamiliar with the notation $\mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbb{I}\right)$, or need a gentle reminder, then this box is for you. The notation $\mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbb{I}\right)$ indicates that the conditional distribution of x_t given x_{t-1} is a multivariate normal distribution with the following characteristics:

1. **Mean**: $\mu = \sqrt{1-\beta_t}x_{t-1}$
2. **Variance-Covariance Matrix**: $\Sigma = \beta_t \mathbb{I}$

As for variance scheduler β_t, keep in mind that it can be learned or fixed. Suppose we have 10 timesteps, if we use a linear variance schedule $\beta_t = \frac{t}{10}$. This means,

- At t = 1, $\beta_1 = 0.1, x_1 \sim \mathcal{N}\left(\sqrt{0.9}x_0, 0.1\mathbb{I}\right)$
- At $t = 2, \beta_2 = 0.2, x_2 \sim \mathcal{N}\left(\sqrt{0.8}x_1, 0.2\mathbb{I}\right)$

- ...
- At t $= 10, \beta_{10} = 1.0, x_{10} = \mathcal{N}(0, \mathbb{I})$

Thus, at the final step the data is completely overwhelmed by noise. To understand this, compare $\mathcal{N}\left(\sqrt{0.8}\vec{x}_1, 0.2\mathbb{I}\right)$ and $\mathcal{N}(0, \mathbb{I})$. The distribution $\mathcal{N}(0, \mathbb{I})$ represents pure noise, with no influence from the original data. Each component is independently distributed with zero mean and unit variance. On the other hand, $\mathcal{N}\left(\sqrt{0.8}x_1, 0.2\mathbb{I}\right)$ represents a mix of the original data x_1 and noise with the mean $\sqrt{0.8}x_1$ indicating that the data still retains some influence from the previous state x_1, but scaled down
The variance defined by $\beta_t\mathbb{I}$ indicates that the covariance matrix is diagonal and that there is no correlation between different components:

$$\beta_t\mathbb{I} = \begin{pmatrix} \beta_t & 0 & \cdots & 0 \\ 0 & \beta_t & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \beta_t \end{pmatrix}$$

What this implies is that the distribution of the noise in different parts of the image is independent of each other.

The complete forward process can be described by the following equation:

$$q\left(x_{T:1} \mid \vec{x}_0\right) = \prod_{i=1}^{T} q\left(x_t \mid x_{t-1}\right) \tag{9.9}$$

For example, for T $= 10$ and $\beta = \frac{t}{10}$:

$$q\left(x_{10:1} \mid x_0\right) = \mathcal{N}\left(\sqrt{0.9}x_0, 0.1\mathbb{I}\right)\mathcal{N}\left(\sqrt{0.8}x_0, 0.2\mathbb{I}\right)\ldots\mathcal{N}(0, \mathbb{I})$$

Thus, the data sample $\vec{x}_0$ gradually loses its distinguishable features as the step t becomes larger. Eventually when T $\to \infty, x_T$ is equivalent to an isotropic Gaussian Distribution.

Thus far, our kernel as given by (9.9) has the following problem: whenever we need a latent sample $\vec{x}_t$, we have to perform $t-1$ steps in the Markov chain. Since this is inefficient (and more importantly we would need the next step for deriving the reverse kernel), we would like to directly go from timestep 0 to timestep t in the process. That is, we would like x_t to be directly conditioned on x_0 rather than intermediate representations x_{t-1}, x_{t-2}... and so on.

To do so, the authors introduce reparameterization trick for $x_t \sim \mathcal{N}\left(\sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbb{I}\right)$:

$$x_t = \mu + \sigma \odot \epsilon_{t-1} \text{ where } \epsilon_{t-1} \sim \mathcal{N}(0, \mathbb{I}) \tag{9.10}$$

Since $\mu = \sqrt{1-\beta_t}\vec{x}_{t-1}$ and $\sigma^2 = \beta_t$, we can write the sampled vector x_t as following:

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_{t-1} \tag{9.11}$$

By introducing the following change in variables:

$$\begin{cases} \alpha_t = 1 - \beta_t \\ \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i \end{cases} \tag{9.12}$$

We can write (9.11) as:

$$\vec{x}_t = \sqrt{\alpha_t}\vec{x}_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1} \tag{9.13}$$

For the time step t − 1, introduce reparameterization again:

$$x_{t-1} = \sqrt{\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_{t-1}}\epsilon_{t-2} \tag{9.14}$$

Plug this,

$$x_t = \sqrt{\alpha_t\alpha_{t-1}} \quad x_{t-2} + \sqrt{\alpha_t\left(1-\alpha_{t-1}\right)}\epsilon_{t-2} + \sqrt{1-\alpha_t}\epsilon_{t-1} \tag{9.15}$$

Since both ϵ_{t-2} and ϵ_{t-1} are Gaussian, we can merge them together where recall that when we merge two Gaussians with different variance $\mathcal{N}\left(0, \sigma_1^2\mathbb{I}\right)$ and $\mathcal{N}\left(0, \sigma_2^2\mathbb{I}\right)$, the new distribution is $\mathcal{N}\left(0, \left(\sigma_1^2 + \sigma_2^2\right)\mathbb{I}\right)$. Thus, define $\bar{\epsilon}_{t-2} = \epsilon_{t-1} + \epsilon_{t-2}$

$$\begin{aligned} x_t &= \sqrt{\alpha_t\alpha_{t-1}} \quad x_{t-2} + \left(\sqrt{\alpha_t\left(1-\alpha_{t-1}\right) + 1 - \alpha_t}\right)\bar{\epsilon}_{t-2} \\ x_t &= \sqrt{\alpha_t\alpha_{t-1}} \quad x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\epsilon}_{t-2} \end{aligned} \tag{9.16}$$

Thus, we see that if we continue this iteratively for t = 0, we will obtain:

$$x_t = \sqrt{\bar{\alpha}_t}\vec{x}_0 + \sqrt{1-\bar{\alpha}_t}\bar{\epsilon}_t \tag{9.17}$$

$$q\left(x_t \mid x_0\right) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, \left(1-\bar{\alpha}_t\right)\mathbb{I}\right) \tag{9.18}$$

Equation (9.18) is thus the forward process that let us go from the original image $\vec{x}_0$ up to complete noise image step by step.

9.2 Reverse diffusion process - Finding a Closed form of Reverse Kernel

If we can reverse the above kernel as encapsulated in (9.9) and sample from $q(x_{t-1} \mid x_t)$, we will be able to recreate the true sample from a Gaussian noise input:

$$q(x_T) \sim \mathcal{N}(0, \mathbb{I}) \tag{9.19}$$

The Markov chain for the reverse diffusion starts from where the forward process ends, i.e., at timestep T, where the data distribution has been converted into (nearly an) isotropic gaussian distribution. The PDF of the reverse diffusion process is an "integral" over all the possible pathways we can take to arrive at a data sample starting from pure noise $\mathbf{x}_T$:

$$p_\theta(x_0) = \int p_\theta(x_{0:T})\, dx_{1:T} \tag{9.20}$$

In terms of a discrete kernel, we have:

$$\begin{cases} p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t) \\ p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \end{cases} \tag{9.21}$$

Where $p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$ highlights our assumption that the reverse kernels should also be gaussians that slowly take us to our original distribution. Furthermore, we denote the mean and covariance of these kernels with $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ respectively. Thus, the mean μ_θ and standard deviation Σ_θ is a function of both x_t, t and is parametrized by θ.

Difficulties of estimating the reverse kernel $q(x_{t-1} \mid x_t)$

We cannot easily estimate $q(x_{t-1} \mid x_t)$ using $p_\theta(x_{t-1} \mid x_t)$ because the estimation of $q(x_{t-1} \mid x_t)$ requires the entire dataset. If that sounds similar to the problem that we encountered with variational autoencoders, then that is because it is. In particular, $q(x_t \mid x_{t-1})$ and $q(x_{t-1} \mid x_t)$ are related by Bayes Theorem:

$$q(x_t \mid x_{t-1}) = \frac{q(x_{t-1} \mid x_t)\, q(x_t)}{q(x_{t-1})} \tag{9.22}$$

It's the intractability of $q(x_{t-1})$, where it is computed over all possible values of x_t, that causes the problem:

$$q(x_{t-1}) = \int q(x_{t-1} \mid x_t)\, q(x_t)\, dx_t \tag{9.23}$$

Due to this intractability of $q(x_{t-1} \mid x_t)$, we need to learn a $p_\theta(x_{t-1} \mid x_t)$ to

approximate these conditional probabilities $q\left(x_{t-1} \mid x_t\right)$ in some other way.

It is (9.21) that our model must learn. In particular, we must learn $\mu_\theta\left(x_t, t\right)$ and $\Sigma_\theta\left(x_t, t\right)$ of $p_\theta\left(x_{t-1} \mid x_t\right)$, which will estimate the true mean $\tilde{\mu}_t$ and $\tilde{\Sigma}$ of $q\left(x_t \mid x_{t-1}\right)$. In the original paper, the authors make a simplifying assumption that the variance $\Sigma_\theta\left(x_t, t\right)$ is not a parameter the network needs to learn, but in later works such as "Improved Denoising Diffusion Probabilistic Models," they also learn the variance.

To begin considering the estimate of $\mu_\theta\left(x_t, t\right)$, let us consider the forward pass $q\left(x_{t-1} \mid x_t\right)$ and invert it using Bayes' theorem:

$$q\left(x_{t-1} \mid x_t\right)=\frac{q\left(x_t \mid x_{t-1}\right) q\left(x_{t-1}\right)}{q\left(x_t\right)} \tag{9.24}$$

Assume now x_{t-1} and x_t are conditioned on x_0 (which they are from equation (9.9) [I told you the step was important for finding a tractable form of reverse kernel]. It follows:

$$q\left(x_{t-1} \mid x_t, x_0\right)=\frac{q\left(x_t \mid x_{t-1}, x_0\right) q\left(x_{t-1} \mid x_0\right)}{q\left(x_t \mid x_0\right)} \tag{9.25}$$

Recall that the Gaussian is of the form $\mathrm{f_x(x)}=\frac{1}{\sqrt{2 \pi \sigma^2}} \exp \left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$. Lets ignore the normalization factors and instead just focus on the exponentials:

- $\mu_1=\sqrt{1-\beta_\mathrm{t}} \mathrm{x}_{\mathrm{t}-1}=\sqrt{\alpha_\mathrm{t}} x_{\mathrm{t}-1}$ and $\sigma_1^2=\beta_\mathrm{t}$ for $q(x_t|x_{t-1}, x_0)$
- $\mu_2=\sqrt{\bar{\alpha}_{\mathrm{t}-1}}$ and $\sigma_2=\sqrt{1-\bar{\alpha}_{t-1}}$ for $q\left(\mathrm{x}_{\mathrm{t}-1} \mid \mathrm{x}_0\right)$
- $\mu_3=\sqrt{\bar{\alpha}_t}$ and $\sigma_3=\sqrt{1-\bar{\alpha}_\mathrm{t}}$ for $q\left(x_\mathrm{t} \mid \mathrm{x}_0\right)$

Plugging these values of mean and standard deviation inside (9.37), we obtain the following:

$$q\left(x_{t-1} \mid x_t, x_0\right) \propto \exp \left(-\frac{1}{2}\left(\frac{\left(x_t-\sqrt{\alpha_t} x_{t-1}\right)^2}{\beta_t}\right)+\frac{\left(x_{t-1}-\sqrt{\alpha_{t-1}} x_0\right)^2}{1-\bar{\alpha}_{t-1}}-\frac{\left(x_t-\sqrt{\bar{\alpha}_t} x_0\right)^2}{1-\bar{\alpha}_t}\right)$$

Forget about the exponential and work on its argument:

$$\frac{x_t^2-2 \sqrt{\alpha_t} x_t x_{t-1}+\alpha_t x_{t-1}^2}{\beta_t}+\frac{x_{t-1}^2-2 \sqrt{\alpha_{t-1}} x_0 x_{t-1}+\bar{\alpha}_{t-1} x_0^2}{1-\bar{\alpha}_{t-1}}+\frac{\left(x_t-\sqrt{\bar{\alpha}_t} x_0\right)^2}{\bar{\alpha}_t}$$

Collecting x_{t-1}^2, x_{t-1} terms together and denoting the other terms as $C\left(x_t, x_0\right)$:

$$\left(\left(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}\right) x_{t-1}^2\right)-\left(\frac{2 \sqrt{\alpha_t}}{\beta_t} x_t+\frac{2 \sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}} x_0\right) x_{t-1}+C\left(x_t, x_0\right) \tag{9.26}$$

Where $C\left(x_t, x_0\right)$ is some function not involving x_{t-1} and therefore is not of interest to us since the mean and standard deviation of the gaussian can be worked out from the terms involving x_{t-1}^2 and x_{t-1} as can be seen from the structure of the equation below:

$$\mathrm{f_x}\left(x_{t-1}\right) \propto \exp\left[-\frac{1}{2}\left(\frac{x_t-\mu}{\sigma}\right)^2\right]$$

$$\mathrm{f_x}\left(x_{t-1}\right) \propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}x_{t-1}^2+\frac{\mu^2}{\sigma^2}+2\frac{\mu}{\sigma^2}x_{t-1}\right)\right] \tag{9.27}$$

Let us denote the inverse standard deviation as $\tilde{\beta}_t = \frac{1}{\sigma^2}$. Comparing (9.27) and (9.26), we see that the deviation assumes the form:

$$\tilde{\beta}_t = 1/\left(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}\right) = 1/\left(\frac{\alpha_t\left(1-\bar{\alpha}_{t-1}\right)+\beta_t}{\beta_t\left(1-\bar{\alpha}_{t-1}\right)}\right)\left[\alpha\bar{\alpha}_{t-1}=\bar{\alpha}_t\right] \tag{9.28}$$

$$\tilde{\beta}_t = 1/\left(\frac{\alpha_t+\bar{\alpha}_t+\beta_t}{\beta_t\left(1-\bar{\alpha}_{t-1}\right)}\right) \tag{9.29}$$

$$\tilde{\beta}_t = \left(\frac{\beta_t\left(1-\bar{\alpha}_{t-1}\right)}{\alpha_t+\bar{\alpha}_t+\beta_t}\right) = \left(\frac{\beta_t\left(1-\bar{\alpha}_{t-1}\right)}{\left(1-\beta_t\right)+\bar{\alpha}_t+\beta_t}\right)\left[\text{ using } \alpha_t = 1-\beta_t\right] \tag{9.30}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{\left(1-\bar{\alpha}_t\right)}\beta_t \tag{9.31}$$

Let us now estimate the mean using the term $x_{t-1}\frac{\mu}{\sigma^2}$ as represented in (9.27) :

$$\begin{aligned}
\tilde{\mu}\left(x_t, x_0\right) &= \left(\frac{\sqrt{\alpha_t}}{\beta_t}x_t+\frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x_0\right)\frac{1-\bar{\alpha}_{t-1}}{\left(1-\bar{\alpha}_t\right)}\beta_t \\
\tilde{\mu}\left(x_t, x_0\right) &= \sqrt{\alpha_t}\frac{1-\bar{\alpha}_{t-1}}{\left(1-\bar{\alpha}_t\right)}x_t+\frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x_0\cdot\frac{1-\bar{\alpha}_{t-1}}{\left(1-\bar{\alpha}_t\right)}\beta_t \\
\tilde{\mu}\left(x_t, x_0\right) &= \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{\left(1-\bar{\alpha}_t\right)}x_t+\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{\left(1-\bar{\alpha}_t\right)}x_0\beta_t
\end{aligned} \tag{9.32}$$

Sinc $x_t = \sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\varepsilon$, it follows that:

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t-\sqrt{1-\bar{\alpha}_t}\varepsilon_t\right) \tag{9.33}$$

Plugging (9.33) into (9.32), we obtain:

$$\tilde{\mu}\left(x_t, x_0\right) = \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t}x_t+\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t-\sqrt{1-\bar{\alpha}_t}\varepsilon_t\right)$$

$$\tilde{\mu}_t\left(x_t, \varepsilon_t\right)=\frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \varepsilon_t\right) \tag{9.34}$$

Equation (9.34) represents the mean of the distribution that our model must learn to estimate.

9.3 The loss function of Diffusion Models

We already talked about the intractability of $q\left(x_{t-1} \mid x_t\right)$. Just like in VAEs, we will use variational inference to mitigate this problem by using a parametrized family of Gaussians $p_\theta\left(x_{1: T} \mid x_0\right)$ to estimate our target distribution $q\left(x_{1: T} \mid x_0\right)$ with KL divergence as the metric of choice. Let $p_\theta\left(x_0\right)$ be our prior, then our ELBO assumes the following form: $L_{\mathrm{ELBO}}=\log p_\theta\left(x_0\right)-D_{K L}\left(q\left(x_{1: T} \mid x_0\right) \| p_\theta\left(x_{1: T} \mid x_0\right)\right)$. To minimize this, we must maximize the following:

$$-\log p_\theta\left(x_0\right) \leq-\log p_\theta\left(x_0\right)+D_{K L}\left(q\left(x_{1: T} \mid x_0\right) \| p_\theta\left(x_{1: T} \mid x_0\right)\right) \tag{9.35}$$

$$=-\log p_\theta\left(x_0\right)+\mathbb{E}_{x_{1: T} \sim q\left(x_{1: T} \mid x_0\right)}\left[\log \frac{q\left(x_{1: T} \mid x_0\right)}{p_\theta\left(x_{1: T} \mid x_0\right)}\right] \tag{9.36}$$

Consider $p_\theta\left(x_{1: T} \mid x_0\right)$. Applying Bayes' rule,

$$p_\theta\left(x_{1: T} \mid x_0\right)=\frac{p_\theta\left(x_0 \mid x_{1: T}\right) p\left(x_{1: T}\right)}{p_\theta\left(x_0\right)}. \tag{9.37}$$

Using the identity $P(A \mid B)=\frac{P(A, B)}{P(B)}$, we can expand (9.37) as:

$$p_\theta\left(x_{1: T} \mid x_0\right)=\frac{p_\theta\left(x_0, x_{1: T}\right)}{p_\theta\left(x_0\right) p\left(x_{1: T}\right)} p\left(x_{1: T}\right)=\frac{p_\theta\left(x_0, x_{1: T}\right)}{p_\theta\left(x_0\right)}=\frac{p_\theta\left(x_{0: T}\right)}{p_\theta\left(x_0\right)} \tag{9.38}$$

Plugging (9.38) back in the logarithm of (9.36):

$$\log \frac{q\left(x_{1: T} \mid x_0\right)}{p_\theta\left(x_{1: T} \mid x_0\right)}=\log \left(\frac{q\left(x_{1: T} \mid x_0\right)}{p_\theta\left(x_{0: T}\right)}\right)+\log \left(p_\theta\left(x_0\right)\right) \tag{9.39}$$

.

Substituting (9.39) into (9.36), it follows:

$$-\log p_\theta\left(x_0\right) \leq \mathbb{E}_{x_{1: T} \sim q\left(x_{1: T} \mid x_0\right)}\left[\log \frac{q\left(x_{1: T} \mid x_0\right)}{p_\theta\left(x_{1: T} \mid x_0\right)}\right]=L_{V L B} \tag{9.40}$$

Equation (9.40) represents a great deal of progress. In fact, it represents our loss function. Let's develop it more:

$$\log \frac{q\left(x_{1: T} \mid x_0\right)}{p_\theta\left(x_{1: T} \mid x_0\right)}=\log \frac{\prod_{t=1}^T q\left(x_t \mid x_{t-1}\right)}{p_\theta\left(x_t\right) \prod_{t=1}^T p_\theta\left(x_{t-1} \mid x_t\right)}$$

$$\log \frac{q\left(x_{1:T} \mid x_0\right)}{p_\theta\left(x_{1:T} \mid x_0\right)} = -\log p_\theta\left(x_t\right) + \sum_{t=1}^{T} \log\left(\frac{q\left(x_t \mid x_{t-1}\right)}{p_\theta\left(x_{t-1} \mid x_t\right)}\right)$$

$$\log \frac{q\left(x_{1:T} \mid x_0\right)}{p_\theta\left(x_{1:T} \mid x_0\right)} = -\log p_\theta\left(x_t\right) + \sum_{t=2}^{T} \log\left(\frac{q\left(x_t \mid x_{t-1}\right)}{p_\theta\left(x_{t-1} \mid x_t\right)}\right) + \log\left(\frac{q\left(x_1 \mid x_0\right)}{p_\theta\left(x_0 \mid x_1\right)}\right). \tag{9.41}$$

Where we moved the first element under the summation outside in the last step. For the numerator in the summation, we can use Bayes' rule where we condition on the initial state:

$$q(x_t|x_{t-1}) = q(x_t|x_{t-1}, x_0) = \frac{q\left(x_{t-1} \mid x_t, x_0\right) q\left(x_t \mid x_0\right)}{q\left(x_{t-1} \mid x_0\right)}. \tag{9.42}$$

Plugging this back into (9.41):

$$-\log p_\theta\left(x_t\right) + \sum_{t=2}^{T} \log\left(\frac{q\left(x_{t-1} \mid x_t, x_0\right) q\left(x_t \mid x_0\right)}{p_\theta\left(x_{t-1} \mid x_t\right) q\left(x_{t-1} \mid x_0\right)}\right) + \log\left(\frac{q\left(x_1 \mid x_0\right)}{p_\theta\left(x_0 \mid x_1\right)}\right). \tag{9.43}$$

Simplifying the summation in (9.43)

We can split the summation in (9.43) as follows:

$$\sum_{t=2}^{T} \log\left(\frac{q\left(x_{t-1} \mid x_t, x_0\right) q\left(x_t \mid x_0\right)}{q\left(x_{t-1} \mid x_0\right)}\right) = \sum_{t=2}^{T} \log\left(\frac{q\left(x_{t-1} \mid x_t, x_0\right)}{p_\theta\left(x_{t-1} \mid x_t\right)}\right) + \sum_{t=2}^{T} \log\left(\frac{q\left(x_t \mid x_0\right)}{q\left(x_{t-1} \mid x_0\right)}\right)$$

Consider the second summation on the right-hand side. The denominator is one step after the numerator. That means if we change the summation of the logs into a multiplication of the arguments in the log, the denominator of the current t will cancel the numerator of $t+1$. So, only the first term in the numerator and the last term in the denominator will survive. Therefore, the summation can be replaced by the term below on the right-hand side:

$$\sum_{t=2}^{T} \log\left(\frac{q\left(x_{t-1} \mid x_t, x_0\right) q\left(x_t \mid x_0\right)}{q\left(x_{t-1} \mid x_0\right)}\right) = \sum_{t=2}^{T} \log\left(\frac{q\left(x_{t-1} \mid x_t, x_0\right)}{p_\theta\left(x_{t-1} \mid x_t\right)}\right) + \log\left(\frac{q\left(x_T \mid x_0\right)}{q\left(x_1 \mid x_0\right)}\right)$$

Plugging the simplified form of the summation into (9.43):

$$
\begin{aligned}
&-\log p_\theta\left(x_t\right)+\sum_{t=2}^T \log\left(\frac{q\left(x_{t-1} \mid x_t, x_0\right)}{p_\theta\left(x_{t-1} \mid x_t\right)}\right)+\log\left(\frac{q\left(x_T \mid x_0\right)}{q\left(x_1 \mid x_0\right)}\right)+\log\left(\frac{q\left(x_1 \mid x_0\right)}{p_\theta\left(x_0 \mid x_1\right)}\right) \\
&-\log p_\theta\left(x_t\right)+\sum_{t=2}^T \log\left(\frac{q\left(x_{t-1} \mid x_t, x_0\right)}{p_\theta\left(x_{t-1} \mid x_t\right)}\right)+\log\left(\frac{q\left(x_T \mid x_0\right)}{q\left(x_1 \mid x_0\right)} \frac{q\left(x_1 \mid x_0\right)}{p_\theta\left(x_0 \mid x_1\right)}\right) \\
&\log q\left(x_T \mid x_0\right)-\log p_\theta\left(x_t\right)+\sum_{t=2}^T \log\left(\frac{q\left(x_{t-1} \mid x_t, x_0\right)}{p_\theta\left(x_{t-1} \mid x_t\right)}\right)-\log p_\theta\left(x_0 \mid x_1\right) \\
&\log\left(\frac{q\left(x_T \mid x_0\right)}{p_\theta\left(x_t\right)}\right)+\sum_{t=2}^T \log\left(\frac{q\left(x_{t-1} \mid x_t, x_0\right)}{p_\theta\left(x_{t-1} \mid x_t\right)}\right)-\log p_\theta\left(x_0 \mid x_1\right) \qquad (9.44)
\end{aligned}
$$

We can represent (9.44) in KL divergence notation:

$$
\begin{aligned}
D_{KL}\left(q\left(x_T \mid x_0\right) \| p_\theta\left(x_T\right)\right)+\sum_{t=2}^T D_{KL}\left(q\left(x_{t-1} \mid x_t, x_0\right) \| p_\theta\left(x_{t-1} \mid x_t\right)\right) \\
-\log p_\theta\left(x_0 \mid x_1\right) \qquad (9.45)
\end{aligned}
$$

If we plug this back into Equation (9.40), our final loss reads:

$$
\begin{aligned}
L_{VAE}=\mathbb{E}_q\Bigg[D_{KL}\left(q\left(x_T \mid x_0\right) \| p_\theta\left(x_T\right)\right)-\log p_\theta(x_0|x_1) \\
+\sum_{t=2}^T D_{KL}\left(q\left(x_{t-1} \mid x_t, x_0\right) \| p_\theta\left(x_{t-1} \mid x_t\right)\right)\Bigg] \qquad (9.46)
\end{aligned}
$$

Let's label each component in the variational lower bound loss separately:

$$
L_{\mathrm{VLB}}=L_T+L_{T-1}+\cdots+L_0 \qquad (9.47)
$$

Where,

$$
\begin{aligned}
L_T &= D_{KL}\left(q\left(\mathbf{x}_T \mid, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_T\right)\right) && (9.48)\\
L_t &= D_{KL}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right) \text{ for } 1 \leq t \leq T-1 && (9.49)\\
L_0 &= -\log p_\theta\left(\mathbf{x}_0 \mid, \mathbf{x}_1\right) && (9.50)
\end{aligned}
$$

The first component on the right-hand side, L_T, can be ignored while training since q has no learnable parameters and x_T is just a Gaussian noise (remember we are at the end of the forward pass). Let's look closer now to the second component L_t. This component calculates the KL divergence between $q(x_t)$ and $p(x_t)$ where,

$$p_\theta\left(x_{t-1} \mid x_t\right)=\mathcal{N}\left(x_{t-1} ; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right) \tag{9.51}$$

$$q\left(x_{t-1} \mid x_t, x_0\right)=\mathcal{N}\left(x_{t-1} ; \tilde{\mu}_t\left(x_t, \varepsilon_t\right), \Sigma_t\left(x_t, t\right)\right) \tag{9.52}$$

Parameterization of L_t for Training Loss: Since we are dealing with Gaussians and the only parameter is the mean, using the KL divergence leads to the following loss:

$$\begin{aligned}
L_t &= \mathbb{E}_{x_0,\varepsilon}\left[\frac{1}{2\sigma_t^2}\left\|\tilde{\mu}_t\left(x_t, \varepsilon_t\right)-\mu_\theta\left(x_t, t\right)\right\|^2\right] \\
L_t &= \mathbb{E}_{x_0,\varepsilon}\left[\frac{1}{2\sigma_t^2}\left\|\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t\right)-\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(x_t, t\right)\right)\right\|^2\right] \\
L_t &= \mathbb{E}_{x_0,\varepsilon}\left[\frac{\left(1-\alpha_t\right)^2}{2\alpha_t\left(1-\bar{\alpha}_t\right)\sigma_t^2}\left\|\epsilon_t-\epsilon_\theta\left(x_t, t\right)\right\|^2\right]
\end{aligned} \tag{9.53}$$

In the DDPM paper, it is found empirically that the training works better if the scaling factor is omitted:

$$\begin{aligned}
L_t &= \mathbb{E}_{t\sim[1,T],\mathbf{x}_0\sim q(\mathbf{x}_0),\varepsilon_t}\left[\left\|\boldsymbol{\epsilon}_t-\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right)\right\|^2\right] \\
L_t &= \mathbb{E}_{t\sim[1,T],\mathbf{x}_0\sim q(\mathbf{x}_0),\varepsilon_t}\left[\left\|\boldsymbol{\epsilon}_t-\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\vec{x}_0+\sqrt{1-\bar{\alpha}_t}\bar{\epsilon}_t, t\right)\right\|^2\right]
\end{aligned} \tag{9.54}$$

10 Score-Based Models

A very natural way to motivate Score-Based models is to consider them in context of Energy-based models (EBM). EBMs are based on the idea of associating a scalar energy value with each configuration of the variables of interest where lower energy corresponds to more likely or more plausible configurations:

$$P_\theta(\vec{x})=\frac{e^{-E_\theta(\vec{x})}}{Z_\theta} \tag{10.1}$$

Where Z_θ is the partition function and normalizes the distribution:

$$Z_\theta=\int d\vec{x} e^{-E_\theta(\vec{x})} \tag{10.2}$$

In MLE, we maximize the probability distribution $\log\left(P_\theta(\vec{x})\right)$:

$$\max_\theta\left[\log\left(P_\theta(\vec{x})\right)\right]=\max_\theta\left[E_\theta(\vec{x})-\log Z_\theta\right] \tag{10.3}$$

Which can be shown to be solving for the following:

$$\nabla_\theta \mathcal{L} = -\nabla_\theta E_\theta(\vec{x}) + \mathbb{E}_{\vec{x}' \sim P_\theta(\vec{x})} \left[\nabla_\theta f_\theta \left(\vec{x}' \right) \right] \tag{10.4}$$

In this formalism, the expectation $\mathbb{E}_{\vec{x}' \sim P_\theta(\vec{x})} \left[\nabla_\theta E_\theta \left(\vec{x}' \right) \right]$ can be challenging due to the intractability of the partition function Z_θ. Since $P_\theta(\vec{x})$ is defined as $-\frac{e^{-f_\theta(\vec{x})}}{Z_\theta}$ and Z_θ involves summation over entire state, direct sampling from $P_\theta(\vec{x})$ is difficult. The expectation in equation is typically approximated using techniques like Markov Chain Monte Carlo (MCMC) or contrastive divergence.

Due to these difficulties in estimating Z_θ, a popular way of bypassing this task in recent times have emerged. This involves solving for $\nabla_x p(x)$ instead. (These are vectors by the way. I haven't bolded or put an arrow above because most of the time I assume it is straightforward to see that we are working with vectors). Notice what happens if we instead try to solve for $\nabla_x p(x)$. The normalization in equation (10.1) disappears. This leads us to define the notion of the score function:

$$\begin{aligned} s_\theta(x) &= \nabla_x \log p_\theta(x) = -\nabla_x E_\theta(x) - \nabla_x \log Z_\theta \\ s_\theta(x) &= \nabla_x \log p_\theta(x) = -\nabla_x E_\theta(x) \end{aligned} \tag{10.5}$$

Where notice that we take the gradient with respect to the data points $\{\vec{x}_1, \ldots, \vec{x}_n\}$ rather than with respect to the parameters $\left\{\vec{\theta}_1, \ldots, \vec{\theta}_n\right\}$. The term $\nabla_x \log Z_\theta$ evaluates to 0 since Z_θ is a constant with respect to x. Since (10.5) is independent of the normalizing constant Z_θ, this significantly expands the family of models that we can tractably use because we don't need any special architectures to make the normalizing constant tractable.

Properties of $s_\theta(x)$

$\nabla_{\vec{x}} \log p_\theta(\vec{x})$ is a vector field that maps from $\mathbb{R}^d \to \mathbb{R}^d$ where d is the dimensionality of the data. Given an input image $\vec{x} = [x_1, x_2, \ldots x_D]$, the score represents the derivative of the scalar-log probability function with respect to each component of the vector $\vec{x}$:

$$\nabla_x \log p_\theta(x) = [\partial_{x_1} \log p_\theta(x), \partial_{x_2} \log p_\theta(x), \ldots, \partial_{x_D} \log p_\theta(x)] \tag{10.6}$$

For example, if $\vec{x}$ is a 2 -dimensional vector $[x_1, x_2]$ and $p_\theta(\vec{x})$ represents a 2D Gaussian distribution. The gradient $\nabla_{\vec{x}} \log p_\theta(\vec{x})$ would be a 2 -dimensional vector that tells you about the direction in which the probability density $p_\theta(\vec{x})$ is increasing. At each point $\vec{x}$, the gradient would point towards the mean of the Gaussian, as that's where the probability is highest. You can see this easily. Take $p_\theta(\vec{x}) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{|\vec{x}-\mu|^2}{2\sigma^2}\right)$ s.t $\log p_\theta(x) = -\frac{|\vec{x}-\mu|^2}{2\sigma^2}$. The derivative being $-\frac{\vec{x}-\mu}{\sigma^2}$.

This is a vector that points from $\vec{x}$ towards the mean μ with magnitude inversely proportional to σ^2

Score function is a conservative vector field if it satisfies Equation (10.5): Recall that a conservative vector field can be expressed as the gradient of the scale potential function. Mathematically, a vector field $\vec{F}$ is conservative if:

$$\vec{F} = -\nabla\phi(x) \tag{10.7}$$

A conservative vector field has the following properties:

- The integral of the vector field between two points does not depend on the path taken.
- The curl of the field is zero: $\nabla \times \vec{F} = 0$

From equation (10.5), we see that $\vec{F} = \log p_\theta(\vec{x})$ satisfies (10.7) with the associated scalar function being $E_\theta(\vec{x})$. This means that the line integral between two points $\int_{x_1}^{x_2} \nabla_{\vec{x}} \log p_t(\vec{x}) d\vec{x}$ is independent of the path taken from $\vec{x}_1$ to $\vec{x}_2$. This is important for MCMC methods. If a score function is conservative, we can sum over random paths, and the total integral will still reflect a true and consistent estimate. Without conservativeness, you'd get different results based on the choice of random paths, undermining the accuracy of the Monte Carlo method.

10.1 Introduction to Score-Matching:

The goal of score-matching is to minimize the difference between p_{data} and p_θ by optimizing the Fisher Divergence. For the sake of simplicity, we consider the 1-D case:

$$\mathbb{E}_{p_{\text{data}}(\vec{x})} \frac{1}{2} \left[\|\nabla_x \log p_{\text{data}}(\vec{x}) - \nabla_x \log p_\theta(\vec{x})\|_2^2 \right] \tag{10.8}$$

Where the subscript 2 in $\|.\|_2$ refers to the standard L^2 norm (also know as the Euclidean norm) such that $\|v\|_2 = (\sum v_i^2)^{\frac{1}{2}}$ and thus $\|.\|_2^2$ indicates L^2 norm squared such that $\|v\|_2^2 = \sum v_i^2$. Further expanding (10.8):

$$\begin{aligned}
&= \frac{1}{2} \int p_{\text{data}}(\vec{x}) \left(\nabla_{\vec{x}} \log p_{\text{data}}(\vec{x}) - \nabla_{\vec{x}} \log p_\theta(\vec{x}) \right)^2 d\vec{x} \\
&= \frac{1}{2} \int p_{\text{data}}(\vec{x}) \left(\nabla_{\vec{x}} \log p_{\text{data}}(\vec{x}) \right)^2 d\vec{x} + \frac{1}{2} \int p_{\text{data}}(\vec{x}) \left(\nabla_{\vec{x}} \log p_\theta(\vec{x}) \right)^2 d\vec{x}
\end{aligned}$$

$$- \int p_{\text{data}}\left(\vec{x}\right) \nabla_{\vec{x}} \log p_{\text{data}}\left(\vec{x}\right) \nabla_{\vec{x}} \log p_\theta(\vec{x}) d\vec{x} \tag{10.9}$$

Note that the first term is simply a constant and can be ignored during optimization because it involves the data distribution $p_{\text{data}}\left(\vec{x}\right)$ and its gradient $\nabla_{\vec{x}} \log p_{\text{data}}\left(\vec{x}\right)$. Since $p_{\text{data}}\left(\vec{x}\right)$ is the true data distribution, it does not depend on the model parameters θ. Therefore, this term is independent of θ and remains constant with respect to θ.

Simplifying the last term in (10.9)

The last term can be written as following:

$$\begin{aligned} \int p_{\text{data}}\left(\vec{x}\right) \nabla_{\vec{x}} \log p_\theta(\vec{x}) \nabla_{\vec{x}} \log p_\theta(\vec{x}) d\vec{x} &= \int p_{\text{data}}\left(\vec{x}\right) \frac{\nabla_{\vec{x}} p_{\text{data}}\left(\vec{x}\right)}{p_{\text{data}}\left(\vec{x}\right)} \nabla_{\vec{x}} \log p_\theta(\vec{x}) d\vec{x} \\ &= \int \nabla_{\vec{x}} \log p_\theta(\vec{x}) \nabla_{\vec{x}} p_{\text{data}}\left(\vec{x}\right) d\vec{x} \end{aligned} \tag{10.10}$$

Applying integration by parts on this::

$$\begin{aligned} &\int \nabla_{\vec{x}} \log p_\theta(\vec{x}) \nabla_{\vec{x}} p_{\text{data}}\left(\vec{x}\right) d\vec{x} = p_{\text{data}}\left(x\right) \nabla_{\vec{x}} \log p_\theta(\vec{x})|_{-\infty}^{\infty} - \int \nabla_{\vec{x}}^2 \log p_\theta(\vec{x}) p_{\text{data}}\left(\vec{x}\right) d\vec{x} \\ &\int \nabla_{\vec{x}} \log p_\theta(\vec{x}) \nabla_{\vec{x}} p_{\text{data}}\left(\vec{x}\right) d\vec{x} = - \int \nabla_{\vec{x}}^2 \log p_\theta(\vec{x}) p_{\text{data}}\left(\vec{x}\right) d\vec{x} \\ &\int \nabla_{\vec{x}} \log p_\theta(\vec{x}) \nabla_{\vec{x}} p_{\text{data}}\left(\vec{x}\right) d\vec{x} = -\mathbb{E}_{p_{\text{data}}}\left[\nabla_{\vec{x}}^2 \log p_\theta(\vec{x})\right] \end{aligned} \tag{10.11}$$

Where p_{data} evaluates to 0 at infinity in order for it to be a valid PMF.

Plugging these all into (10.9), we obtain:

$$\begin{aligned} &= \text{const} \ + \frac{1}{2} \mathbb{E}_{p_{\text{data}}\left(\vec{x}\right)}\left[\left(\nabla_x \log p_\theta(\vec{x})\right)^2\right] + \mathbb{E}_{p_{\text{data}}\left(\vec{x}\right)}\left[\nabla_x^2 \log p_\theta(\vec{x})\right] \\ &= \mathbb{E}_{p_{\text{data}}\left(\vec{x}\right)}\left[\nabla_x^2 \log p_\theta(\vec{x}) + \frac{1}{2}\left(\nabla_x \log p_\theta(\vec{x})\right)^2\right] + \ \text{const} \end{aligned} \tag{10.12}$$

We can easily extend this into a multidimensional context, the result of which is:

$$\mathbb{E}_{p_{\text{data}}\left(\vec{x}\right)}\left[\text{tr}\left(\nabla_x^2 \log p_\theta(\vec{x})\right) + \frac{1}{2}\left\|\nabla_x \log p_\theta(\vec{x})\right\|_2^2\right] + \ \text{const} \tag{10.13}$$

In terms of the score function $s_\theta(\vec{x})$, one can write equation 10.13 as following:

$$\mathbb{E}_{p_{\text{data}}\left(\vec{x}\right)}\left[\text{tr}\left(\nabla_x^2 s_\theta(\vec{x}),\right) + \frac{1}{2}\left\|s_\theta(\vec{x})\right\|_2^2\right] \tag{10.14}$$

We are specifically interested in instances where f_θ is parametrized as a neural network. Recall from equation (10.5) that:

$$s_\theta(x) = \nabla_{\vec{x}} \log p_\theta(\vec{x}) = -\nabla_{\vec{x}} E_\theta(\vec{x}) \tag{10.15}$$

Therefore, we can rewrite the score-matching objective as:

$$\mathbb{E}_{p_{\text{data}}(\vec{x})}\left[\text{tr}\left(\nabla_x^2 E_\theta(\vec{x})\right) + \frac{1}{2}\left\|\nabla_x E_\theta(\vec{x})\right\|_2^2\right] + \text{ const} \tag{10.16}$$

Sliced-Score Matching:: In equation (10.16), while the first-order gradient can be simply obtained via backpropagation, $\nabla_x^2 E_\theta(\vec{x})$ is very computationally costly. To circumvent this problem, the authors propose random projection which reduces dimensionality of data down to scalars. Quoting Yang Song:

"We propose sliced score matching to greatly scale up the computation of score matching. The motivating idea is that one dimensional data distribution is much easier to estimate for score matching. We propose to project the scores onto random directions, such that the vector fields of scores of the data and model distribution become scalar fields. We then compare the scalar fields to determine how far the model distribution is from the data distribution. It is clear to see that the two vector fields are equivalent if and only if their scalar fields corresponding to projections onto all directions are the same."

Equation (10.13) gave the standard version of fisher divergence. The random projection version of Fisher divergence is

$$L\left(\vec{\theta}, p_{\vec{v}}\right) = \frac{1}{2}\mathbb{E}_{p_{\vec{v}}}\mathbb{E}_{p_{\text{data}}(\vec{x})}\left[\left(\vec{v}^T\nabla_x \log p_{\text{data}}(\vec{x}) - \vec{v}^T\nabla_x \log p_\theta(\vec{x})\right)^2\right] \tag{10.17}$$

Expanding the squared terms:

$$L\left(\vec{\theta}, p_{\vec{v}}\right) = \frac{1}{2}\mathbb{E}_{p_{\vec{v}}}\mathbb{E}_{p_{\text{data}}(\vec{x})}\left[\left(\vec{v}^T s_\theta(\vec{x})\right)^2 + \left(\vec{v}^T s_{\text{data}}(\vec{x})\right)^2 - 2\left(\vec{v}^T s_\theta(\vec{x})\right)\left(\vec{v}^T s_{\text{data}}(\vec{x})\right)\right]$$

Where $\left(\vec{v}^T s_{\text{data}}(\vec{x})\right)^2$ is a constant with respect to θ again such that:

$$L\left(\vec{\theta}, p_{\vec{v}}\right) = \frac{1}{2}\mathbb{E}_{p_{\vec{v}}}\mathbb{E}_{p_{\text{data}}(\vec{x})}\left[\frac{1}{2}\left(\vec{v}^T s_\theta(\vec{x})\right)^2 - \left(\vec{v}^T s_\theta(\vec{x})\right)\left(\vec{v}^T s_{\text{data}}(\vec{x})\right)\right] + C \tag{10.18}$$

Simplifying the last term in (10.18)

Just like we did before, lets works on the last term:

$$-\mathbb{E}_{p_{\vec{v}}} \int \left(\vec{v}^T s_\theta\right) \left(\vec{v}^T s_{\text{data}}\right) p_{\text{data}}\, d\vec{x} = -\mathbb{E}_{p_{\vec{v}}} \int \left(\vec{v}^T s_\theta\right) \left(\vec{v}^T \frac{\nabla_{\bar{x}} p_{\text{data}}}{p_{\text{data}}}\right) p_{\text{data}}\, d\vec{x}$$
$$= -\mathbb{E}_{p_{\vec{v}}} \int \left(\vec{v}^T s_\theta(\vec{x})\right) \left(\vec{v}^T \nabla_{\vec{x}} p_{\text{data}}\left(\vec{x}\right)\right) d\vec{x}$$

We can express the dot product over the summation over individual components $\vec{v}^T \nabla_{\vec{x}} p_{\text{data}}\left(\vec{x}\right) = \sum_i v_i \frac{\partial p_{\text{data}}\left(x\right)}{\partial x_i}$ such that the above can be written as:

$$-\mathbb{E}_{p_{\vec{v}}} \int \left(\vec{v}^T s_\theta(\vec{x})\right) \left(\vec{v}^T \nabla_{\vec{x}} p_{\text{data}}\left(\vec{x}\right)\right) d\vec{x} = -\mathbb{E}_{p_{\vec{v}}} \sum_i \int \left(\vec{v}^T s_\theta(\vec{x})\right) \left(v_i \frac{\partial p_{\text{data}}\left(x\right)}{\partial x_i}\right) d\vec{x}$$
$$= \mathbb{E}_{p_{\vec{v}}} \int \left(\vec{v}^T s_\theta(\vec{x})\right) \left(\vec{v} \cdot p_{\text{data}}\left(\vec{x}\right)\right) d\vec{x}$$
$$= \mathbb{E}_{p_{\vec{v}}} \mathbb{E}_{p_{\text{data}}\left(\vec{x}\right)} \left[\vec{v}^T s_\theta(\vec{x}) \vec{v}\right] \tag{10.19}$$

In the last part, we used integration by parts just like we did for naïve score matching. Thus, our loss function reads:

$$L\left(\vec{\theta}, p_{\vec{v}}\right) = \frac{1}{2} \mathbb{E}_{p_{\vec{v}}} \mathbb{E}_{p_{\text{data}}\left(\vec{x}\right)} \left[\frac{1}{2} \left(\vec{v}^T s_\theta(\vec{x})\right)^2 + \vec{v}^T s_\theta(\vec{x}) \vec{v}\right] \tag{10.20}$$

Intuitively, the equation forces the two distributions to get closer according to some random projection $\vec{v}$. Since the projection is random, there exists a guarantee that optimizing this quantity will bring p_θ closer to the real data distribution. We then have the following unbiased estimator:

$$J_{N,M}\left(\theta; p_v\right) = \frac{1}{N} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} \left[\vec{v}_{ij}^T \nabla_x \log p_\theta\left(\vec{x}_i\right) \vec{v}_{ij} + \left(\frac{1}{2} s_\theta\left(\vec{x}_i\right)\right)^2\right] \tag{10.21}$$

NSCN: Score-based generative modeling with multiple noise perturbations: There is a problem with naïve score matching as encapsulated by (10.12) equation. The key challenge is the fact that the estimated score functions are inaccurate in low density regions, where few data points are available for computing the score matching objective. This is expected because the l_2 differences are weighted by $p(x)$ and are largely ignored where $p(x)$ is small. This behavior can lead to subpar results. ?

How can we bypass the difficulty of accurate score estimation in regions of low data density? The solution that Yang Song came upon was to perturb data points with noise and train score-based models on the noisy data points instead. When the noise magnitude is sufficiently large, it can populate low data density regions to improve the accuracy of estimated scores.

Another question remains: how do we choose an appropriate noise scale for the perturbation process? Larger noise can obviously cover more low density regions for better score estimation, but it over-corrupts the data and alters it significantly from the original distribution. Smaller noise, on the other hand, causes less corruption of the original data distribution, but does not cover the low density regions as well as we would like.

To achieve the best of both worlds, they use multiple scales of noise perturbations simultaneously .Suppose we always perturb the data with isotropic Gaussian noise, and let there be a total of L
increasing standard deviations $\sigma_1 < \sigma_2 < \cdots \sigma_L$. We first perturb the data distribution $p(\vec{x})$ with each of the Gaussian noise $\mathcal{N}\left(0, \sigma_i^2 \mathbb{I}\right), \mathrm{i} = 1, 2, \ldots$ L to obtain a noise-perturbed distribution:

$$p_{\sigma_i}(\vec{x}) = \int p_{\text{data}}\ (\vec{x}')\, \mathcal{N}\left(\vec{x}; \vec{x}', \sigma_i^2 \mathbb{I}\right) d\vec{x}' \tag{10.22}$$

Note that we can easily draw samples from $p_{\sigma_i}(\vec{x})$ by sampling $\vec{x} \sim p(\vec{x})$ and computing $\vec{x} + \sigma_i \vec{z}$ with $\vec{z} \sim \mathcal{N}(0, \mathbb{I})$

Next, we estimate the score function of each noise-perturbed distribution, $\nabla_{\vec{x}} \log p_{\sigma_i}(\vec{x})$ by training a Noise-conditioned Score-Based model $s_\theta(\vec{x}, i)$ (also called a Noise Conditional Score Network, or NCSN when parametrized with a neural network) with a score matching, such that $s_\theta(\vec{x}, i) \approx \nabla_{\vec{x}} \log p_{\sigma_i}(\vec{x})$ for all $\mathrm{i} = 1, 2, \ldots,$ L

We apply multiple scales of Gaussian noise to perturb the data distribution (first row), and jointly estimate the score functions for all of them (second row).

The training objective for $s_\theta(\vec{x}, i)$ is a weighted sum of Fisher divergences for all noise scales. In particular, we use the objective below:

$$\sum_{i=1}^{L} \lambda(i) \mathbb{E}_{p_{\sigma_i}(\vec{x})} \left[\left\| \nabla_{\vec{x}} \log p_{\sigma_i}(\vec{x}) - s_\theta(\vec{x}, i) \right\|_2^2 \right] \tag{10.23}$$

Where $\lambda(i) \in \mathbb{R}_{>0}$ is a positive weight function, often chosen to be $\lambda(i) = \sigma_i^2$ where now the unperturbed data distribution $p_{\text{data}}\left(\vec{x}\right)$ has been replaced with the noise induced distribuition $p_{\sigma_i}(\vec{x})$.

11 Limitations and Open Challenges

Discussion of current limitations and potential challenges for future research in these domains.

A Appendices

A.1 Code Resources

A.1.1 Code for Euler-Maruyama Simulation

```
mu = 0.1
sigma = 0.2
X0 = 1.0
T = 1.0
N = 1000
dt = T / N
t = np.linspace(0, T, N + 1)

X = np.zeros(N + 1)
X[0] = X0
for i in range(1, N + 1):
    dW = np.random.normal(0, np.sqrt(dt))
    X[i] = X[i - 1] + mu * X[i - 1] * dt + sigma * X[i - 1] *
       dW


true_X = X0 * np.exp((mu - 0.5 * sigma**2) * t + sigma * np.
   cumsum(np.random.normal(0, np.sqrt(dt), N + 1)))

plt.figure(figsize=(10, 6))
plt.plot(t, X, label='Simulated GBM ( E u l e r Maruyama )',
   linewidth=2)
plt.plot(t, true_X, label='True GBM Path', linestyle='dashed',
   linewidth=2)
plt.title('Simulated vs True Path of Geometric Brownian Motion'
   )
plt.xlabel('Time')
plt.ylabel('X(t)')
plt.grid(True)
plt.legend()
plt.show()
```

A.2 Suplementary Material

Supplemental proofs and derivations supporting the discussed methodologies.

A.2.1 Sum of Gaussian Variables

Claim: Let X and Y be independent random variables that are normally distributed (and therefore also jointly so). Then their sum is also normally distributed. That is, if $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$, and $Z = X + Y$, then $Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

Proof: We give the elementary proof using the moment-generating function (MGF). The MGF of a normal random variable $X \sim N(\mu_X, \sigma_X^2)$ is given by:

$$M_X(t) = \mathbb{E}[e^{tX}] = \exp\left(\mu_X t + \frac{1}{2}\sigma_X^2 t^2\right) \tag{A.1}$$

Similarly, for $Y \sim N(\mu_Y, \sigma_Y^2)$:

$$M_Y(t) = \mathbb{E}[e^{tY}] = \exp\left(\mu_Y t + \frac{1}{2}\sigma_Y^2 t^2\right) \tag{A.2}$$

If X and Y are independent, the MGF of their sum $Z = X + Y$ is the product of their MGFs:

$$\begin{aligned} M_Z(t) &= \mathbb{E}[e^{tZ}] = \mathbb{E}[e^{tX}] \cdot \mathbb{E}[e^{tY}] \\ &= \exp\left(\mu_X t + \frac{1}{2}\sigma_X^2 t^2\right) \cdot \exp\left(\mu_Y t + \frac{1}{2}\sigma_Y^2 t^2\right) \end{aligned} \tag{A.3}$$

Combining these:

$$\begin{aligned} M_Z(t) &= \exp\left(\mu_X t + \frac{1}{2}\sigma_X^2 t^2 + \mu_Y t + \frac{1}{2}\sigma_Y^2 t^2\right) \\ &= \exp\left((\mu_X + \mu_Y)t + \frac{1}{2}(\sigma_X^2 + \sigma_Y^2)t^2\right) \end{aligned} \tag{A.4}$$

The MGF $M_Z(t) = \exp\left((\mu_X + \mu_Y)t + \frac{1}{2}(\sigma_X^2 + \sigma_Y^2)t^2\right)$ is the MGF of a normal distribution with mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$.

Thus, we conclude that:

$$Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2). \tag{A.5}$$

A.2.2 Change of Variable Theorem

Let X be a continuous random variable with probability density function $f_X(x)$, and let $Y = g(X)$, where $g : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable and monotonic function. Then the PDF of Y, denoted by $f_Y(y)$, is given by:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|,$$

where g^{-1} is the inverse function of g, and y belongs to the range of $g(X)$.

We show that the above holds by the assumption of *conservation of probability*. When we transform X to Y, the probability mass must remain the same. Consider a small intervals around X and Y:

- A small interval $[x, x + dx]$ around X has a probability $f_X(x) \cdot dx$.
- The corresponding interval around Y is $[y, y + dy]$ with the probability $f_Y(x) \cdot du$

The probability in both intervals should be the same, which translates to:

$$f_Y(y) \cdot dy = f_X(x) \cdot dx \tag{A.6}$$

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| \tag{A.7}$$

Where $x = g^{-1}(y)$ such that:

$$f_Y(y) = f_X(x) \left| \frac{d}{dy} g^{-1}(y) \right|. \tag{A.8}$$

For a transformation involving multiple variables, the theorem generalizes using the Jacobian matrix. Let X be a random vector (just think of it as a collection of random variables clumped together in a single vector), and let $Y = g(X)$ be a transformation where g is a differentiable function. The goal is to find the joint PDF $f_Y(y)$. The Jacobian matrix J of the transformation g captures how the transformation changes the differential volume elements:

$$J = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \tag{A.9}$$

The determinant of the Jacobian matrix represents the factor by which the volume element is scaled during the transformation. For the transformation to preserve probability measures, the differential volume elements must scale correctly. Thus, the change of variable formula for the joint PDF in higher dimensions is:

$$f_Y(y) = f_X\big(g^{-1}(y)\big) \left| \det \left(\frac{\partial g^{-1}(y)}{\partial y} \right) \right| \tag{A.10}$$

A.3 Additional Figures

Figures and diagrams for extended illustration of key concepts.

Acknowledgments

This work has been supported and guided by Dr. Muhammad Faryad. His mentorship and expertise have been instrumental in developing this research.

References